

# Methods of Inference for Nonparametric Curves and Surfaces

Mitchum T. Bock

*A Dissertation Submitted to the  
University of Glasgow  
for the degree of  
Doctor of Philosophy*

Department of Statistics

December 1999

©Mitchum T. Bock

ProQuest Number: 13834242

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 13834242

Published by ProQuest LLC (2019). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code  
Microform Edition © ProQuest LLC.

ProQuest LLC.  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106 – 1346

Department of Statistics,  
University of Glasgow,  
Mathematics Building,  
University Gardens,  
Glasgow G12 8QW

GLASGOW  
UNIVERSITY  
LIBRARY

M754

(copy 2)

*To Mum, Dad and Tania*



# Abstract

Nonparametric regression models offer attractive extensions to the familiar approaches of parametric regression. They adapt to departures from standard parametric forms and therefore have the potential to capture features which may otherwise go unnoticed. This property accounts for the large volume of work in the area of estimation of nonparametric models which has emerged over the last two decades.

Inferential techniques using nonparametric model fits, however, have not been so quick to develop. This thesis contributes to this area of research by examining the task of assessing covariate effects via comparisons of nonparametric model fits. In particular, the asymptotic and finite sample bias properties of estimates obtained via local linear smoothers are a major consideration and methods of inference which take into account these properties are developed.

Chapter 1 introduces and presents an overview of existing methods of estimation and inference amongst nonparametric regression. Chapter 2 focuses on the task of inference by considering the estimation of the error variance in the nonparametric model context. Special attention is given to the development and assessment of difference based estimators in the presence of two covariates. It is shown that difference based estimators are a viable alternative, in terms of accuracy, to standard residual based estimators.

Chapters 3 and 4 employ the estimators of Chapter 2 in the development of test procedures which make comparisons amongst a class of bivariate nonparametric regression models. Chapter 3 develops the theoretical properties of several forms of the test statistic, with particular attention given to statistics based on direct comparisons of fitted values. The theory also highlights the rôle of centred smoothers and equivalent degrees of smoothing when nonparametric model fits are compared.

The simulation studies reported in Chapter 4 compare the novel approaches developed in Chapter 3 with standard approaches based on differences in residual

sums of squares, i.e. approximate F-tests. The results show that direct comparisons of fitted values offer an improvement in some settings and never perform less favourably in others. The choice of the error variance estimator is shown to be crucial, with different design spaces requiring different estimators. Specific attention is also given to the effect of correlation amongst the covariates on the tests' performances. Chapter 4 closes with an application of the methods to a real data set describing the spatial distribution of sea bed fauna in the Great Barrier Reef.

Chapter 5 extends these methods beyond models with two covariates to models with an unlimited number of additive linear terms and a nonparametric component involving at most two covariates. Recent results which derive the asymptotic properties of models of this form show that the favourable properties of local linear regression in the bivariate setting extend to this multidimensional setting. Results of a simulation study are reported and show that there is much to be gained by making a direct comparison of fitted values in conjunction with a careful choice of the estimator of error variance.

Chapters 6 and 7 describe applied projects in environmental and medical contexts respectively. Both of the sets of data contain relationships amongst covariates which are best described using nonparametric models. Chapter 6 considers 14 years of water quality monitoring data from the Firth of Clyde, Scotland. Interest lies in describing relationships between pollutants and environmental factors, including long term trends and seasonal patterns. Chapter 7 investigates the relationship between short term dosage of an immunosuppressive drug and the long term outcome of kidney transplantation patients.

Chapter 8 concludes with a summary of the main findings of the thesis and a discussion of potential future work in this area. Although progress has been made in the settings considered in the thesis, further extensions are required before nonparametric modelling will achieve its full potential.

# Acknowledgements

It has been a great privilege to be the first Australian to benefit from the Scots Australian Scholarship scheme and I am indebted to the Scots Australian Council for their vision and Clydesdale Bank for their sponsorship. Special thanks goes to Sir Iain Noble, Douglas Briggs, David Bruce and Zena Sime for their assistance.

Without the encouragement and support of my supervisor Professor Adrian Bowman this thesis would not exist. It has been an honour to serve my “research apprenticeship” with access to the wisdom and insight of one so accomplished yet humble. Adrian’s commitment, extending far beyond the realms of academia, has been an inspiration and I trust that my future path will reflect his excellent example.

Collaborations with Drs. Brian Miller and Stuart Roger on applied projects proved very valuable experiences and I thank them for providing access to their fields of research. I have also benefited through interaction with members of the Department of Statistics. The Department certainly lives up to Glasgow’s reputation of friendliness and this has been nowhere more obvious than amongst the postgrads with whom I had the pleasure of “serving time”. I wish you all much success and fulfilment in your future years.

Though family and friends of the Australian variety have been distant, I have been helped by their confidence in me and expressions of love. I hope that my infrequent and sparse contact will not be seen as a gauge of my appreciation. I have also been blessed through the friendships that have grown over the last three years here in Glasgow. A great debt of gratitude is owed to the quality men and women I have come to know and respect, particularly through my involvement with The Navigators and St Silas Episcopal Church. As examples of such, I would like to thank Donald MacFarlane and Tim McKenzie for their patience and love as they bore the brunt of the “final year”.

And lastly, “thanks be to God, who in Christ always leads us in triumphal procession, and through us spreads in every place the fragrance that comes from knowing him” (II Cor 2:14).

# Contents

<b>1</b>	<b>Introduction and Overview</b>	<b>1</b>
1.1	Introduction . . . . .	1
1.2	Univariate Nonparametric Regression . . . . .	2
1.2.1	Estimation . . . . .	2
1.2.2	Smoothing parameter selection . . . . .	9
1.2.3	Methods of inference . . . . .	13
1.3	Multivariate Nonparametric Regression . . . . .	18
1.3.1	Generalisations to a multidimensional setting . . . . .	18
1.3.2	Estimation . . . . .	22
1.3.3	Smoothing parameter selection . . . . .	24
1.3.4	Methods of inference . . . . .	24
1.4	Software . . . . .	26
1.5	Overview of Thesis . . . . .	26
<b>2</b>	<b>Error Variance Estimation</b>	<b>28</b>
2.1	The Rôle of the Error Variance . . . . .	28
2.2	Univariate Estimators . . . . .	30
2.2.1	Residual sum of squares estimators . . . . .	30
2.2.2	Difference based estimators . . . . .	32
2.3	Variance Estimation in Higher Dimensions . . . . .	35
2.3.1	RSS based approaches . . . . .	36
2.3.2	Difference based approaches . . . . .	37
2.4	Improved RSS Based Estimator of $\sigma^2$ . . . . .	39
2.4.1	Undersmoothing . . . . .	39

2.4.2	Double smoothing . . . . .	40
2.5	Defining a Difference Based Estimator in Two Dimensions . . . . .	42
2.6	Properties of Estimators . . . . .	49
2.6.1	Finite sample bias and variance . . . . .	50
2.6.2	Asymptotic properties . . . . .	50
2.7	Simulation Study . . . . .	51
2.7.1	Estimators under investigation . . . . .	51
2.7.2	Simulated conditions . . . . .	52
2.7.3	Simulation results . . . . .	55
2.8	Discussion . . . . .	60
<b>3</b>	<b>Bivariate Model Inference: Theory</b>	<b>70</b>
3.1	Introduction . . . . .	70
3.2	Models Under Consideration . . . . .	71
3.2.1	Defining the fitted values and bias of each of the models . . . . .	71
3.2.2	Constant (no-effect) model . . . . .	74
3.2.3	Simple and bivariate linear regression model . . . . .	74
3.2.4	Univariate smooth model . . . . .	75
3.2.5	Semiparametric model . . . . .	77
3.2.6	Additive model . . . . .	80
3.2.7	Bivariate smooth model . . . . .	82
3.2.8	Summary . . . . .	84
3.3	Comparing Models via Fitted Values . . . . .	86
3.3.1	Comparing linear fits . . . . .	86
3.3.2	Nonparametric case . . . . .	88
3.4	Distributional Properties of CFV . . . . .	90
3.5	Tests Under Investigation . . . . .	92
3.5.1	Approximate F-test . . . . .	93
3.5.2	Two-moment corrected F-test . . . . .	95
3.5.3	Quadratic form approach . . . . .	97
3.5.4	Summary . . . . .	99

<b>4</b>	<b>Bivariate Model Inference: Simulations</b>	<b>100</b>
4.1	Introduction . . . . .	100
4.2	Model Comparisons Simulation Study: Regular Grid Design . . .	102
4.2.1	Regular grid: $\sigma^2$ known . . . . .	102
4.2.2	Regular grid: $\sigma^2$ estimated . . . . .	106
4.3	Model Comparison Simulation Study: Random Design . . . . .	115
4.3.1	Random design: $\sigma^2$ known . . . . .	125
4.3.2	Random design: $\sigma^2$ estimated . . . . .	132
4.3.3	Random design: correlated covariates . . . . .	138
4.4	Model Comparisons Applied to the Reef Data . . . . .	142
4.5	Discussion . . . . .	149
<b>5</b>	<b>Inference &amp; Semiparametric Models</b>	<b>155</b>
5.1	Introduction: Models Under Consideration . . . . .	155
5.2	Defining the Fitted Values and Bias . . . . .	158
5.2.1	Univariate SAM . . . . .	162
5.2.2	Additive SAM . . . . .	163
5.2.3	Bivariate SAM . . . . .	166
5.2.4	Summary . . . . .	167
5.3	Model Comparisons . . . . .	168
5.4	Simulation Study . . . . .	170
5.4.1	Regular grid designs . . . . .	170
5.4.2	Random designs . . . . .	174
5.5	Discussion . . . . .	178
<b>6</b>	<b>Firth of Clyde Analysis</b>	<b>181</b>
6.1	Introduction & Background . . . . .	181
6.2	The Data . . . . .	182
6.2.1	The variables . . . . .	182
6.2.2	Sampling frequency . . . . .	184
6.3	Analyses & Results . . . . .	185
6.3.1	The distributions of measured variables . . . . .	185
6.3.2	Changes in the concentrations of variables with depth . . .	187

6.3.3	Evidence for the presence of long term trends . . . . .	189
6.3.4	Seasonal patterns . . . . .	190
6.3.5	Important aspects of the dynamics of the water system . .	193
6.4	Discussion Points . . . . .	195
6.4.1	Relationships between variables . . . . .	195
6.4.2	Potential for eutrophication . . . . .	197
6.4.3	Dynamics of the water system . . . . .	197
6.4.4	Effects of stratification . . . . .	198
6.5	Conclusions . . . . .	198
<b>7</b>	<b>Cyclosporin Treatment Analysis</b>	<b>200</b>
7.1	Introduction and Background . . . . .	200
7.2	Previous and Current Work . . . . .	202
7.3	Study Design . . . . .	204
7.3.1	Selecting a cohort of patients . . . . .	204
7.3.2	Defining and determining outcome . . . . .	205
7.3.3	Demographic and clinical variables . . . . .	207
7.4	Cyclosporin Variables . . . . .	209
7.4.1	Dosage data . . . . .	210
7.4.2	Blood levels data . . . . .	212
7.4.3	Summary of cyclosporin variables . . . . .	215
7.5	Analysis and Results . . . . .	215
7.5.1	Univariate analysis . . . . .	216
7.5.2	Multivariate analysis . . . . .	219
7.6	Discussion . . . . .	221
<b>8</b>	<b>Conclusions</b>	<b>224</b>
8.1	Introduction . . . . .	224
8.2	Estimation of Error Variance . . . . .	225
8.2.1	Summary and discussion . . . . .	225
8.2.2	Future work . . . . .	226
8.3	Inference Amongst Bivariate Models . . . . .	226
8.3.1	Summary and discussion . . . . .	226

8.3.2	Future work . . . . .	229
8.4	Inference with Broader Classes of Models . . . . .	231
8.4.1	Summary and discussion . . . . .	231
8.4.2	Future work . . . . .	232
<b>A</b>	<b>Software</b>	<b>235</b>
<b>B</b>	<b>Simulation Results</b>	<b>237</b>
	<b>References</b>	<b>256</b>



# List of Tables

3.1	Seven models defining the bivariate class . . . . .	72
3.2	Comparisons amongst the bivariate class of models considered in the simulation studies of Chapter 4 . . . . .	74
3.3	Asymptotic bias expressions of ‘full’ model fits when the regression function is of a ‘reduced’ form . . . . .	85
3.4	Asymptotic bias expressions of semiparametric and additive model fits when the regression function is a univariate function over depen- dent covariates. . . . .	86
3.5	Components of tests to compare model fits . . . . .	99
4.1	Summary of simulations: regular grid, $\sigma^2$ known . . . . .	103
4.2	Summary of simulations: regular grid, $\sigma^2$ estimated . . . . .	108
4.3	Power results: regular grids, $\sigma^2$ estimated . . . . .	116
4.4	Summary of simulations: random designs, $\sigma^2$ known . . . . .	127
4.5	Summary of simulations: random design, $\sigma^2$ estimated . . . . .	133
4.6	Summary of simulations: random design-correlated covariates, $\sigma^2$ estimated . . . . .	139
5.1	Summary of simulations: SAMs . . . . .	173
5.2	Power results: regular grids . . . . .	175
5.3	Power results: random designs. . . . .	178
6.1	List of variables available for each water sample . . . . .	184
6.2	Frequency of sampling at the ‘CMT7’ station . . . . .	185

7.1	Number of patients corresponding to each level of the <i>outcome</i> variable . . . . .	207
7.2	Continuous cyclosporin covariates together with their descriptive statistics. . . . .	215
7.3	Binary cyclosporin covariates indicating the monotonicity of doses. . . . .	215
7.4	Results of log-rank tests examining the effects of categorical covariates on long term kidney outcome . . . . .	218
7.5	Continuous covariate results from univariate proportional hazard model fits . . . . .	219
7.6	Details of the 'baseline' fit . . . . .	220
7.7	Assessing the effect of cyclosporin measures after allowing for baseline influences . . . . .	221
B.1	Size results: regular grid, $\sigma^2$ known . . . . .	238
B.2	Power results: regular grid, $\sigma^2$ known . . . . .	239
B.3	Settings of simulation study over regular grids . . . . .	240
B.4	Size results: regular grid, $\sigma^2$ estimated, corrected F distribution . . . . .	241
B.5	Size results: regular grid, $\sigma^2$ estimated, QF distribution . . . . .	242
B.6	Size results: regular grid, $\sigma^2$ estimated, F distribution . . . . .	243
B.7	Settings of simulation study over random designs, $\sigma^2$ known . . . . .	244
B.8	Size results: uniform design, $\sigma^2$ known . . . . .	245
B.9	Size results: bivariate normal ( $\rho = 0$ ) design, $\sigma^2$ known . . . . .	246
B.10	Size results: bivariate normal ( $\rho = 0.5$ ) design, $\sigma^2$ known . . . . .	247
B.11	Settings of simulation study over random designs, $\sigma^2$ estimated . . . . .	248
B.12	Size results: uniform design ( $n = 100$ ), $\sigma^2$ estimated, QF distribution . . . . .	249
B.13	Size results: uniform design ( $n = 100$ ), $\sigma^2$ estimated, F distribution . . . . .	250
B.14	Size results: uniform design ( $n = 100$ ), $\sigma^2$ estimated, corrected F distribution . . . . .	251
B.15	Size results: uniform design ( $n = 49$ ), $\sigma^2$ estimated, QF distribution . . . . .	252
B.16	Size results: uniform design ( $n = 49$ ), $\sigma^2$ estimated, F distribution . . . . .	253
B.17	Size results: uniform design ( $n = 49$ ), $\sigma^2$ estimated, corrected F distribution . . . . .	254
B.18	Power results: uniform design, $\sigma^2$ estimated, corrected F distribution . . . . .	255

# List of Figures

2.1	Representation of a segment of a rectangular grid showing an interior point ( $\square$ ) with 8 neighbouring points ( $\times$ ). The four triplets, used to define a pseudoresidual at the interior point, are shown using different line styles. . . . .	44
2.2	Realisation of a random design showing the Delaunay triangulation which define neighbouring triplets. . . . .	45
2.3	Realisation of design points with the weights applied to the five triplet distances used to define the pseudoresidual at the centre point	49
2.4	True underlying functions used to simulate data. . . . .	54
2.5	Boxplots showing the distribution of estimates of $\sigma^2$ (solid horizontal line) from different estimators. 500 simulations of data generated over a regular grid were used. . . . .	57
2.6	Boxplots showing the distribution of estimates of $\sigma^2$ from different estimators under different simulation conditions. 500 simulations of data generated over a regular grid were used in each boxplot. .	58
2.7	Plots showing the finite sample biases of different estimators of $\sigma^2$ under different regression functions and sample sizes over a regular grid. . . . .	59
2.8	Boxplots showing the distribution of estimates of $\sigma^2$ (solid horizontal line) from different estimators. 500 simulations of data generated over uniform-random designs were used. . . . .	61
2.9	Boxplots showing the distribution of estimates of $\sigma^2$ (solid horizontal line) from different estimators. 500 simulations of data generated over binorm-random ( $\rho = 0$ ) designs were used. . . . .	62

2.10	Boxplots showing the distribution of estimates of $\sigma^2$ (solid horizontal line) from different estimators. 500 simulations of data generated over binorm-random ( $\rho = 0.5$ ) designs were used. . . . .	63
2.11	Plots showing the finite sample biases of different estimators of $\sigma^2$ under different regression functions over random designs. . . . .	64
3.1	Diagram showing the hierarchy of models and natural model comparisons amongst the bivariate class of models. . . . .	73
4.1	Histograms of p-values of different tests . . . . .	102
4.2	Empirical sizes of tests of model comparisons using known $\sigma^2$ . Design points form a regular square grid. The results listed are the proportion of 1000 simulated p-values less than 0.05 under the reduced model. The solid horizontal lines (when shown) indicate a 99% confidence interval from Bin(1000,0,05). . . . .	105
4.3	Power results: regular grid, $\sigma^2$ known . . . . .	107
4.4	Size results of tests for model comparisons using 4 estimates of $\sigma^2$ . 500 sets of observations from a <i>st.line</i> model were generated over a regular square grid. This (true) model fit was compared with a <i>semi.par</i> fit (linear in X) using 2 test statistics, and p-values calculated using 3 reference distributions, yielding the empirical sizes plotted. . . . .	109
4.5	Size results of tests for model comparisons using 4 estimates of $\sigma^2$ . 500 sets of observations from a <i>1d.sm</i> model were generated over a regular square grid. This (true) model fit was compared with a <i>2d.am</i> fit using 2 test statistics, and p-values calculated using 3 reference distributions, yielding the empirical sizes plotted. . . . .	110
4.6	Size results of tests for model comparisons using 4 estimates of $\sigma^2$ . 500 sets of observations from a <i>1d.sm</i> model were generated over a regular square grid. This (true) model fit was compared with a <i>semi.par</i> fit (linear in Z) using 2 test statistics, and p-values calculated using 3 reference distributions, yielding the empirical sizes plotted. . . . .	111

4.7	Size results of tests for model comparisons using 4 estimates of $\sigma^2$ . 500 sets of observations from a <i>semi.par</i> model were generated over a regular square grid. This (true) model fit was compared with a <i>2d.am</i> fit using 2 test statistics, and p-values calculated using 3 reference distributions, yielding the empirical sizes plotted. . . . .	112
4.8	Size results of tests for model comparisons using 4 estimates of $\sigma^2$ . 500 sets of observations from a <i>2d.am</i> model were generated over a regular square grid. This (true) model fit was compared with a <i>2d.sm</i> fit using 2 test statistics, and p-values calculated using 3 reference distributions, yielding the empirical sizes plotted. . . . .	113
4.9	Three covariate design spaces. . . . .	118
4.10	Expectations of fitted values under different nonparametric models for the <i>regular grid</i> (first panel) and the <i>irregular grid</i> (second panel). In each panel the true model is displayed as a solid line and the expectation of the <i>univariate smooth</i> is displayed as a dotted line. The expected values of the <i>bivariate smooth</i> are shown as $\times$ and the expected values of a <i>two dimensional additive model</i> are shown as $+$ . . . . .	119
4.11	Expectations of fitted values under different nonparametric models for the <i>random</i> design. In each panel the true model is displayed as a solid line and the expectation of the <i>univariate smooth</i> is displayed as a dotted line. The expected values of the <i>bivariate smooth</i> are shown in the first panel as $\times$ and the expected values of a <i>two dimensional additive model</i> are shown in the second panel as $+$ . . . . .	120
4.12	Expectations of fitted values under different nonparametric models for the <i>random</i> design using 625 simulated points. In each panel the true model is displayed as a solid line and the expectation of the <i>univariate smooth</i> is displayed as a dotted line. The expected values of the <i>bivariate smooth</i> are shown in the first panel as $\times$ and the expected values of a <i>two dimensional additive model</i> are shown in the second panel as $+$ . . . . .	123

4.13	Comparisons of Model Biases (CMB) averaged over 500 simulated designs ( $n = 100$ ) from bivariate normal distributions with varying levels of correlation. . . . .	126
4.14	Empirical sizes of tests of model comparisons using known $\sigma^2$ . Design points are realisations of <i>uniform random</i> designs. The results listed are the proportion of 500 simulated p-values less than 0.05 under the reduced model. The solid horizontal lines (when shown) indicate a 99% confidence interval from Bin(500,0,05). . .	129
4.15	Empirical sizes of tests of model comparisons using known $\sigma^2$ . Design points are realisations of <i>bivariate normal</i> designs with ( $\rho = 0$ ). The results listed are the proportion of 500 simulated p-values less than 0.05 under the reduced model. The solid horizontal lines (when shown) indicate a 99% confidence interval from Bin(500,0,05). . . . .	130
4.16	Empirical sizes of tests of model comparisons using known $\sigma^2$ . Design points are realisations of <i>bivariate normal</i> designs with ( $\rho = 0.5$ ). The results listed are the proportion of 500 simulated p-values less than 0.05 under the reduced model. The solid horizontal lines (when shown) indicate a 99% confidence interval from Bin(500,0,05). . . . .	131
4.17	Size traces: random design, $\sigma^2$ estimated, $n = 100$ . . . . .	135
4.18	Size traces: random design, $\sigma^2$ estimated, $n = 49$ . . . . .	136
4.19	Power results: random design, $\sigma^2$ estimated . . . . .	137
4.20	Size traces: random design-correlated covariates, $\sigma^2$ estimated . .	141
4.21	Plots showing the location of sampling points and the spatial distributions of catch scores from one year of the reef data. . . . .	143
4.22	Significance traces testing a univariate smooth fit in longitude against an additive fit for each combinations of model comparison, variance estimator and reference distribution. . . . .	144
4.23	Significance traces of tests based on CFV and corrected F distribution for comparisons amongst semiparametric, additive and bivariate smooth fits to the reef data. . . . .	145

4.24	Bootstrap p-values for the 1d.sm vs. 2d.am model comparison applied to the reef data. . . . .	146
4.25	Significance traces of tests using the gam() and lo() functions in S-Plus for comparisons amongst univariate, semiparametric and additive model fits to the reef data. . . . .	148
4.26	Size traces of four tests to compare a univariate smooth and a bivariate additive fit. Tests '3' and '4' use centred smoothers throughout, tests '1' and '2' use uncentred smoothers in the univariate smooth fit. . . . .	150
4.27	Size traces of four tests to compare a univariate smooth and a bivariate additive fit. Tests '3' and '4' use equivalent smoothing parameters in the two fits. Tests '1' and '2' use a range of smoothing parameters in the additive fit but keep the smoothing parameters in the univariate smooth fit constant at 0.15. . . . .	151
5.1	Empirical sizes, each based on 500 simulated data sets over regular grid designs . . . . .	171
5.2	Empirical sizes, each based on 500 simulated data sets over regular grid designs . . . . .	172
5.3	Empirical sizes, each based on 250 simulated data sets over uniform-random designs. . . . .	176
6.1	Map showing the location of the CMT7 sampling station . . . . .	183
6.2	Density estimates of selected variables both before and after log transformations. Also shown are normal reference bands and tick marks indicating individual values. . . . .	186
6.3	Distributions of selected variables with depth. . . . .	188
6.4	Concentrations across time for dissolved oxygen, nitrate and silicate, each at three depths. . . . .	191
6.5	Estimates of seasonal patterns, derived from pooled data, for selected parameters. . . . .	192
6.6	Scatterplots of relationships between selected variables, at depth of 7m. . . . .	194

6.7	Estimates of seasonal patterns, derived from pooled data, for salinity and temperature. . . . .	196
7.1	Traces of the cyclosporin weight adjusted dose data for six patients. Also shown are the measures of magnitude and variation over the two years and the second year only. . . . .	211
7.2	Traces of the cyclosporin levels data for six patients with an estimate of the underlying trend using local linear regression (smoothing parameter of 60 days). An estimate of the average level over the second year is shown together with $\pm$ the estimate of the standard deviation of the levels around the underlying trend. . . . .	212
7.3	Kaplan-Meier estimates of the survival function estimate at the different levels of the categorical factors based on the 206 'raw dose' data. . . . .	217



# Chapter 1

## Introduction and Overview of Nonparametric Regression

### 1.1 Introduction

Statistical models seek to describe and explore the relationships between observable phenomena. Regression models, a large and important class of statistical models, propose some underlying relationship, a *regression function*, which relates the mean value of a *response* variable to the values of a set of *predictor* variables. Within this class there are two subclasses: *parametric* and *nonparametric* regression models. These differ in the form the regression function is allowed to take. As the name implies, *parametric* models assume the regression function can be expressed as a function with a finite number of (unknown) parameters, whereas *nonparametric* models do not specify a rigid form, rather they allow the regression function to be any *smooth* function of the predictor variables.

This thesis examines and develops methods of inference that apply to a class of nonparametric regression models. Although methods of estimation for these models have received considerable attention in the research literature in recent years, methods of inference have not been considered in similar depth. Before the merits of this modelling approach can be truly appreciated and utilised, a fuller understanding of methods of inference in this context is necessary.

This chapter introduces estimation and existing approaches to inference in the

context of nonparametric modelling. Numerous papers and a number of books have been written on the topic, including some comprehensive and influential treatments by Hastie and Tibshirani (1990), Green and Silverman (1994), Fan and Gijbels (1996) and Simonoff (1996). These sources and many specialised papers have been surveyed and some key results are summarised here.

We shall start by considering the ‘nuts and bolts’ of nonparametric modelling, namely univariate smoothing techniques. These methods yield estimates of underlying smooth regression functions in one covariate. Several techniques are described which represent quite different approaches and have different properties. A brief comparison of two of these techniques, the cubic smoothing spline and local linear smooth, is presented and justification given for the choice of the local linear method as the approach taken in the remainder of the thesis.

Methods for modelling several covariates in a nonparametric manner are introduced in Section 1.3. The natural step of constructing a smoother in more than one dimension is considered. However, issues related to the *curse of dimensionality* arise in this multivariate setting and lead to the investigation of dimensionality reduction techniques. Attention will be focused here on *additive models* since these are utilised extensively throughout the thesis. Existing methods of inference in both the univariate and multivariate setting are described together with some of the challenges associated with them.

Section 1.4 describes the computer software used throughout the thesis to implement the methods developed and perform simulation studies. The chapter closes with an overview of the thesis.

## 1.2 Univariate Nonparametric Regression

### 1.2.1 Estimation

The context in which univariate regression models apply is where bivariate data  $(X_1, Y_1), \dots, (X_n, Y_n)$  are observed and the relationship between pairs of  $X$  and  $Y$  is described by  $Y = m(X) + \varepsilon$  where  $\varepsilon \sim N(0, \sigma^2)$  and  $m(x_0) = E(Y|X = x_0)$  is the *regression function*. It is this function (and sometimes its derivatives) which is estimated when  $Y$  ‘is modelled on’  $X$ .

A parametric approach to the estimation of  $m(\cdot)$  assumes a parametric form and then estimates the unknown parameters. Simple linear regression is the most familiar and widely used regression model. It assumes a linear regression function,  $m(x_0) = \beta_0 + \beta_1 x_0$ , and fits the model by estimating the parameters  $\beta_0$  and  $\beta_1$ . Many extensions and generalisations of the linear regression model have been developed, to the point where classes of models such as Generalised Linear Models form a comprehensive and well understood approach to regression modelling.

The limitations of a parametric approach are clear, however. If the true regression function is far from the imposed parametric form then the modelling bias will be considerable. The consequence of this misspecification is that the estimated regression function is worse than useless, since it could badly misrepresent the true structure. In other words, there is always the risk that the parametric form does not adequately capture all the features of the data and may suggest features which aren't present.

Diagnostic tools have been developed (see Cook and Weisberg 1982) to aid in the assessment of these vital assumptions. These techniques may indeed lead to the transformations of variables or the modification of the parametric model through the introduction of higher order terms. However, there will be situations in which this approach also fails to capture the true form of the regression function. Therefore, an alternate approach is not to impose a rigid functional form on the regression function, but to allow the data to determine the nature of the functional form directly.

Nonparametric regression models, as the name implies, replace the rigid assumption that the regression function corresponds to a specific parametric form with the weaker assumption that the regression function is *smooth*. This natural and often realistic assumption opens the door to the modelling of the dependence of  $Y$  on  $X$  via the following argument:

'If  $m(\cdot)$  is believed to be smooth, then the observations at  $X_i$  near  $x$  should contain information about the value of  $m$  at  $x(\cdot)$ . Thus it should be possible to use something like a local average of the data near  $x$  to construct an estimator of  $m(x)$ '.

This quotation from Eubank (1988 pp.7) states the rationale underpinning the nonparametric approach to model fitting. It is an approach which has been around for some time (Whittaker's 1923 paper is one of the earliest works cited). Indeed, as Fan and Gijbels (1996) point out (pp. 2), this idea even forms a part of the parametric approach, since a first step in checking the assumptions of a parametric model is to view a scatterplot of response ( $Y$ ) against predictor ( $X$ ) during which our eyes look for the smooth function underlying pattern behind the points thereby using the data to suggest the form of the parametric model to be fitted.

Techniques which estimate  $m(\cdot)$  using this philosophy are known as *univariate smoothers* and the estimated (or *fitted*) function they yield are called *smooths*. In order to formalise these ideas of applying a smoother to a scatterplot of points, two issues must be addressed:

- ◊ How should the observations at  $X_i$  near  $x$  be combined (i.e. “averaged”) to obtain an estimate of the regression function at  $x$ ?
- ◊ Which  $X_i$  should be considered “near”  $x$ , i.e. from how large a neighbourhood around  $x$  should observations be taken?

The first issue is the subject of the remainder of this section. The second issue is discussed in Section 1.2.2 since issues of the size of the neighbourhood, the choice of smoothing parameter and the degree of smoothing are synonymous.

Let the local average be expressed in the following notation:

$$\hat{m}(x) = \sum_{i=1}^n w_{xi} Y_i$$

where  $[w_{xi}]_{i=1}^n$  is a sequence of weights different for each value of  $x$ . The question is then what weights to use. Different answers to this question define the different types of smoothers.

### Local polynomial smoothers

The smoother used throughout this thesis belongs to the class of *local polynomial regression models*. This approach is a natural extension of the familiar polynomial

(parametric) regression, since it estimates  $m(x_0)$  by fitting a weighted polynomial regression model to the observations in a local ‘neighbourhood’ surrounding  $x_0$ . Taylor’s expansion for the unknown regression function at the point  $x_0$  in the neighbourhood of  $x$  yields:

$$m(x_0) \approx \sum_{j=0}^p \frac{m^{(j)}(x_0)}{j!} (x_0 - x)^j \equiv \sum_{j=0}^p \beta_j (x_0 - x)^j.$$

This suggests that the regression function  $m(\cdot)$  can be estimated by repeatedly fitting local polynomial models at points in the domain of  $m(\cdot)$ .

In particular, the *local linear regression* estimator of  $m(x_0)$  proceeds by fitting a weighted linear regression locally around  $x_0$ , i.e. find  $\hat{\beta}_0$  and  $\hat{\beta}_1$  which minimise

$$\sum_{i=1}^n (Y_i - (\beta_0 + \beta_1(X_i - x_0)))^2 K_h(X_i - x_0),$$

where  $K_h(\cdot)$  is a kernel weight function defined as  $K_h(\cdot) = K(\cdot/h)/h$  where  $K(\cdot)$  is a symmetric (around zero) probability density  $K(\cdot)$  and  $h$  is the *bandwidth* or *smoothing parameter* which controls the size of the local neighbourhood. Again, Taylor’s expansion indicates that an estimator of  $m(x_0)$  is  $\hat{m}(x_0) = \hat{\beta}_0$ , which can be explicitly defined as

$$\hat{m}(x_0) = \sum_{i=1}^n w_i Y_i / \sum_{i=1}^n w_i, \quad w_i = K_h(X_i - x_0) (S_{n,2} - (X_i - x_0) S_{n,1}),$$

where  $S_{n,j} = \sum_{i=1}^n K_h(X_i - x_0) (X_i - x_0)^j$ .

It will be shown in the following chapters how the properties of the local linear smoother suggest ways in which methods of inference may proceed. For this reason it will be used throughout this work. For a kernel function, a standard Gaussian probability density will be used since its unboundedness ensures finite conditional and unconditional variance of the estimators (see Simonoff, 1996, pp141). It has been observed, however, that: ‘the choice of the kernel function  $K(\cdot)$  is not very important for the performance of the resulting estimators, both theoretically and empirically’ (Fan and Gijbels, 1996, pp. 76). The question of which smoothing parameter to employ is a complex one which will be considered

in Section 1.2.2.

Despite the exclusive adoption of the local linear smoother here, it should be noted that this is only one of a number of widely used nonparametric regression estimators. *Kernel smoothers* are one such alternative to local linear regression. They correspond to the local polynomials of degree zero, i.e. local constant fits. As such they have a particularly simple form; the Nadaraya-Watson estimate (see Nadaraya (1964) and Watson (1964)), for instance, is defined as:

$$\hat{m}(x_0) = \frac{\sum_{i=1}^n K_h(X_i - x_0) Y_i}{\sum_{i=1}^n K_h(X_i - x_0)}.$$

Another popular kernel smoother was proposed by Gasser and Müller (1979). Assuming that the data have already been sorted into increasing order over the  $X$ -variable, the estimate is defined as

$$\hat{m}(x_0) = \sum_{i=1}^n Y_i \int_{s_{i-1}}^{s_i} K_h(u - x_0) du$$

with  $s_i = (X_i + X_{i+1})/2$ ,  $X_0 = -\infty$  and  $X_{n+1} = +\infty$ . Although this approach was derived in the context of equispaced designs, it can also be applied to non-equispaced designs. Its advantage over the Nadaraya-Watson estimator is that it does not require a normalising denominator calculated over the entire design space, since the weights sum to one. This leads to advantages in deriving the asymptotic theory of the smoother.

These definitions make it clear that the local linear and kernel smoothers (and indeed all the local polynomial smoothers) can be written as weighted arithmetic means of the observed  $Y$ 's, that is, they are all *linear smoothers*. Because  $m(\cdot)$  is a function it needs to be estimated over a set of values in its domain. This set of *evaluation points* usually corresponds to the observed predictor variables, although this need not be the case. The estimates at these evaluation points can be conveniently defined as:

$$\hat{\mathbf{m}} = \mathbf{S}\mathbf{y}$$

where  $\hat{\mathbf{m}}$  denotes a vector of fitted values,  $\mathbf{y}$  is the vector of observed values and  $\mathbf{S}$  is a *smoothing matrix* whose rows comprise the weights corresponding to each evaluation point ( $[w_{xi}]_{i=1}^n$  in the notation above).  $\mathbf{S}$  is analogous to the hat matrix in linear regression, although its properties will clearly differ.

It follows that many of the properties of the smoothing techniques can be expressed and understood in terms of the properties of the corresponding smoothing matrices. This property unites the approaches since it introduces a common notation and leads to some theoretical results which are common to all linear smoothers. Indeed at the core of this thesis is a methodology which capitalises on the linear form and the associated properties of the local linear smoother.

Although they share similar forms and origins, the local linear and kernel smoothers do differ with respect to important properties. Accuracy and stability are two important considerations about any estimate. A consideration of the differences in these properties will illuminate the choice of the local linear approach in the work which follows.

Table 1.2.1 (taken from Fan (1992)) lists expressions for the first order asymptotic bias and variance at an interior point of the support of the density of the predictor  $X$  (denote this *design density* by  $f(x)$ ) for each of the three local smoothers described above. Clearly these results favour the local linear method since it shares the better properties of the two kernel methods.

Smoother	Bias	Variance
Nadaraya-Watson	$(m''(x_0) + \frac{2m'(x_0)f'(x_0)}{f(x_0)})b_n$	$V_n$
Gasser-Müller	$m''(x_0)b_n$	$1.5V_n$
Local linear	$m''(x_0)b_n$	$V_n$

$$b_n = -\frac{h^2}{2} \int_{-\infty}^{\infty} u^2 K(u) du \text{ and } V_n = \sigma^2(x_0) / [f(x_0)nh \int_{-\infty}^{\infty} K^2(u) du]$$

The differences are due to the way each method adapts to different designs which is reflected in the way weights are assigned to nearby points. Around points away from the boundary, the Nadaraya-Watson estimate assigns symmetric weights which can lead to large bias, whereas the Gasser-Müller estimate assigns weights which reflect the design density in that region and thus may appear almost random, which induces greater variance. The local linear fit borrows from each approach and adapts automatically to this random design by assigning a

smooth asymmetric weighting scheme. Similarly at the boundaries, the local linear fit adapts readily whereas the two kernel estimators do not. Fan and Gijbels (1996) Section 3.2 gives a thorough overview of the merits of the local polynomial approach.

Another attractive property of the local linear estimate is that its finite sample bias is zero when the true regression function is itself linear (Ruppert and Wand (1994)). This is in contrast to the kernel methods which do not possess this finite sample property (see Fan and Gijbels (1996) pp. 63). This is particularly useful in the context of the model comparisons to be considered in Chapter 3 and further recommends the adoption of local linear smoothing for model inference.

One final difference, which has been highlighted above, is the derivation of the smoothing procedure. A local linear approach has an intuitive appeal since it is a relaxation of the (very) familiar linear regression approach. Indeed, its scope of fits range from the straight line fit to an interpolation of the data. This feature also yields theoretical advantages since at the core of the estimation procedure lies well developed and understood least squares theory. This leads to the simple expressions for local bias and variance which are especially important when univariate smooths are used as the building blocks of higher dimension models such as those introduced in Section 1.3.1.

### Spline smoothers

There also exists, however, another popular smoother which, unlike the kernel and the local polynomial approaches, is not motivated or defined using 'local averages'. Smoothing splines are, rather, defined as solutions to a global optimisation problem. Namely, the smooth is defined as the function  $\hat{m}(\cdot)$  which minimises the penalised sum of squares:

$$PSS(m) = \sum_{i=1}^n (Y_i - g(t_i))^2 + \lambda \int_a^b (m''(x))^2 dx,$$

where  $\lambda$  acts as a smoothing parameter.

The two components of  $PSS(m)$  reflect the two conflicting objectives of the smoothing process, the 'goodness-of-fit' and the 'roughness' of the estimate.



These two properties are traded off, with the size of  $\lambda$  controlling the degree of influence given to each of the two terms. The solution to this minimisation problem belongs to the class of smooths known as natural cubic splines and its existence is unique (Green and Silverman (1994) pp. 17-19).

Like the kernel and local linear methods, the spline approach *does* yields a linear smoother. The explicit definitions of the weights, however, are quite involved (see Section 2.1.2 of Green and Silverman (1994)) and computationally intensive. In practice, algorithms which bypass the calculation of weights are used to obtain the estimated smooth (Reinsch (1967)). The complexity of defining and obtaining the weights has the consequence that the properties of the smoothing spline are less transparent than other methods.

Intuitively, therefore, it is difficult to fathom what a smoothing spline is actually doing with local observations. On a theoretical level the consequence is that expressions for the pointwise bias and variance of the smooth are less straight forward than in the kernel and local linear case and thus less open to comparisons. Approximations to the mean and variance can be derived by recognising that the spline approach is basically a local kernel average with a variable bandwidth, as demonstrated by Silverman (1984). These approximations aren't asymptotically exact, however, and the expression can be complicated. Since the methods of inference which are developed herein are motivated largely by comparisons of asymptotic and exact bias, this property of smoothing splines excludes them from use in this work.

Another property of local polynomial fits which recommends their use in this context is the ease with which they generalise to estimate higher dimensional surfaces. This generalisation is described in Section 1.3.2.

## 1.2.2 Smoothing parameter selection

Having considered the question of which method of smoothing to use, the question remains: *what level of smoothing is suitable?* This issue was phrased in terms of defining neighbourhoods in which to take local averages in Section 1.2.1. The same issue can be described using terminology such as *model complexity*, *degrees of freedom*, *equivalent number of parameters*. Regardless of the terminology, the

issue always reduces to choosing a smoothing parameter.

Common to each of the smoothing procedures described above is the existence of a nonnegative parameter ( $h$  and  $\lambda$  used above, exclusively  $h$  below) which scales the degree of smoothness of the estimator. At the extreme value of zero there is no smoothness imposed and the result is an estimate which interpolates the observed data, thus producing an estimate of high variability. At the other extreme ( $\infty$ ) smoothness is 'maximised'. For instance, the limit of a local-linear smooth (as  $h \rightarrow \infty$ ) is the linear least squares fit to the data, which can possess considerable bias. Clearly some value in the range  $(0, \infty)$  must be used which yields a suitable compromise between variability and bias.

One starting point would be to consider a measure which combines the mean and variance of the estimator. The *mean squared error* at each point  $E\{\hat{m}(x) - m(x)\}^2$  is an obvious choice since it is the sum of the squared bias and variance of the estimator. A natural extension of this to the global and asymptotic realm is the *mean integrated squared error*,

$$MISE(h) = \int E\{\hat{m}_h(x) - m(x)\}^2 f(x) dx,$$

where  $f(x)$  represents the density of the design points. The notation makes it clear that MISE is a function of the smoothing parameter  $h$  and thus the value which minimises MISE,  $h_{MISE}$  say, could be a good choice for the smoothing parameter. There are two general approaches to this optimisation problem.

The first involves deriving expressions for the asymptotic bias and variance of a particular smoother (if they exist). They can be used to approximate the MISE and the minimisation can be performed to yield an expression for  $h_{MISE}$ . Not unexpectedly, the expression for  $h_{MISE}$  will involve unknown quantities, for instance in the local linear case  $h_{MISE}$  involves the design density  $f(\cdot)$ , the conditional variance  $\sigma^2$  and the second derivative of the unknown curve  $m''(\cdot)$  (See Fan and Gijbels (1996) 3.2.3).

These 'plug in' methods address the problem of unknowns via a direct estimation of these quantities which are then substituted ('plugged') into the optimal expression above. This approach can be quite involved since the quantities are often complex and it tends to be very smoother specific (see, for instance, Gasser

*et al.* (1991) for the Gasser-Müller smooth and Ruppert *et al.* (1995) for the local linear case). Fan and Gijbels (1996, Section 4.2) describe an extension of this idea which involves global polynomial (i.e. parametric) fits to estimate the unknown properties, which they term ‘rule of thumb’ methods.

The other approach is to use a ‘classical’ bandwidth selector which minimises an estimate of MISE over  $h$ . One example is a ‘cross-validation’ score, defined as

$$CV(h) = \frac{1}{n} \sum_{i=1}^n \{y_i - \hat{m}_{-i}(x_i)\}^2,$$

where  $y_i - \hat{m}_{-i}(x_i)$  is a measure of how well a smooth of all the data except  $(x_i, y_i)$  (denoted by  $\hat{m}_{-i}(\cdot)$ ) predicts the observed value at  $x_i$ . It can be shown that

$$E\{CV(h)\} = \frac{1}{n} \sum_{i=1}^n E\{\hat{m}_{-i}(x_i) - m(x_i)\}^2 + \sigma^2.$$

Since  $CV(h)$  provides a simple estimator for  $MISE(h)$  (plus a constant) a desirable  $h$  can be defined as that which minimises  $CV(h)$ . The *Generalized Cross Validation* method was developed in the context of smoothing splines by Wahba (1978) and Craven and Wahba (1979) and though its original motivation no longer exists, it continues to receive attention for its theoretical and computational properties which can outperform ordinary cross validation.

Though each of these *data driven* methods for selecting the bandwidth have sound theoretical underpinnings, in practice their performance can be disappointing. Hastie and Tibshirani (1990, Section 3.4.5) perform a small simulation study which demonstrates that the cross-validation approaches can be highly variable and quite misleading. The authors suggest more subjective approaches such as graphical methods and a consideration of the degrees of freedom of the smooth which is equivalent to a smoothing parameter on a more meaningful scale.

A recent proposal by Hurvich *et al.* (1998) modifies the ‘classical’ approach known as the Akaike Information Criteria (AIC). Their claim is that while the attractive generality of the classical approach is retained, their modification doesn’t exhibit the high variability in the choice of smoothing nor the tendency to undersmooth.

Although bandwidth selection is a crucial consideration in its own right, the objective of the smoothing exercise should also be kept in mind. If the objective is estimation, i.e. an accurate description of the underlying regression function, then the choice of smoothing parameter is pivotal and every step should be taken to optimise the properties of the estimate. However, if the smoothing is a means to an inferential end, for instance to investigate the effect of a predictor on the response, then the choice of the smoothing parameter may not be so crucial.

Noting this, several authors have proposed a partial means of circumventing bandwidth selection in contexts of inference. King *et al.* (1991) proposed that p-values of tests using smooths be computed at several different choices of the smoothing parameter. Azzalini and Bowman (1993) independently proposed this idea and termed the plot of the p-values versus the bandwidths a *significance trace*. Provided the trace lies strictly above or below the specified level of significance the tricky question of a suitable bandwidth has been avoided (for this task at least). Clearly ambiguity arises when the trace crosses the significance level. In such a case, however, a single p-value corresponding to a specified smoothing parameter would most probably be close to the significance level and therefore would also be inconclusive.

It has been necessary to adopt this approach in the work which follows. One reason is the scope of the thesis which is primarily focused on the methods of inference amongst nonparametric models and not the question of smoothing parameter selection in the context of model inference. A more fundamental reason is that the methods themselves *require* that automatic *data driven* smoothing parameters be excluded. The reason is the methods are based on certain properties of *linear smoothers*, as indeed all of the methods described above are. However, if the smoothing parameter is a function of the responses then the properties of linear smoothers no longer hold. That is, if  $h$  is random through its dependency of  $Y$  then  $w_{xi}$  are also random (since the  $w_{xi}$  are functions of  $h$ ) and thus the smooth  $\hat{m}(x) = \sum_{i=1}^n w_{xi} Y_i$  is no longer linear in  $Y$ .

Therefore ‘fixed’ bandwidths are used throughout this thesis. A modification of a significance trace is used in some of the simulations to investigate the influence of the smoothing parameter on the size of the test. That is, simulations are

repeated at several levels of the smoothing parameter and plots are made of the empirical sizes (at the  $\alpha = 0.05$  level) of the tests against the smoothing parameter values. We call these *size traces*.

### 1.2.3 Methods of inference

Section 1.2.1 described the main approaches to the estimation of the regression function in the context of univariate nonparametric regression. In this section attention will turn to issues of inference, thereby introducing the focus of the remainder of the thesis. As noted above, the topic of inference amongst nonparametric regression models has not received widespread attention. This is somewhat surprising given that the uses of nonparametric models are severely limited without methods for assessing and comparing the model fits.

For inference to proceed another estimate is required, namely that of the variance of observations around the (unknown) regression function, i.e.  $\sigma^2$ . The estimation of  $\sigma^2$  is fundamental to all methods of inference since it enters any consideration of the accuracy of the regression estimate. Chapter 2 is dedicated to the topic of error variance estimation in the bivariate setting and gives an overview of approaches in the univariate setting. Its importance is noted here by way of introduction, before techniques employing such estimates are described.

But what is meant by ‘methods of inference’ in nonparametric regression? In the univariate case, inferential techniques investigate the form of the unknown regression function  $m(\cdot)$ . Two natural simplifications of the general ‘smooth’ form supposed by a nonparametric model are the *no-effect* model  $m(x) = \beta_0$  and the linear model  $m(x) = \beta_0 + \beta_1 x$ . Comparisons between nonparametric fits applied to different data are also of interest, for instance in an Analysis of Covariance where the covariate effect is permitted to be nonlinear. Another comparison which is sometimes made is between two nonparametric models with different smoothing parameters. This is related to smoothing parameter selection which, as noted above, is not a focus of this current work.

A popular approach to model inference in a variety of settings is via *bootstrap* methods. These computer-intensive approaches overcome the lack of distributional theory by simulating empirical distributions which (hopefully) mirror the

(unknown) sampling distributions of various estimators. Methods have been proposed in a variety of contexts: the selection of a suitable smoothing parameter (Hall, 1990); the construction of confidence intervals for the unknown regression function (Hall, 1992); comparisons for checking parametric fits via a comparison with nonparametric ones (Azzalini et al., 1989) and comparisons of nonparametric curves (King *et al.* 1991).

However, the same difficulty which arises in other approaches to inference also presents itself when bootstrap methods are employed, namely *bias*. As Davison and Hinkley (1997) note, ‘unfortunately the inherent bias of most nonparametric regression methods distorts the fitted values and the residuals, and thence biases the resampling scheme’. This comment is made in the context of constructing confidence limits but it also applies when nonparametric fits are compared. As such, bootstrap methods do not provide a comprehensive solution to problems of inference.

Let us therefore review some of the methods of inference which have been suggested to compare fitted models. In both types of comparisons, nonparametric vs. parametric and nonparametric vs. nonparametric, there are two general approaches: confidence bands and global tests.

### Confidence bands

A confidence interval is an informative addition to any point estimate, providing information on the accuracy and consistency of an estimator. In the context of estimating an unknown regression function, this idea leads to the calculation of a confidence band to accompany the estimate of the regression function. This can either comprise pointwise intervals or, for the purposes of model inference, can try and take into account the global nature of the regression function estimate, and thus provide a confidence region for the entire function. As with other tools of inference, however, methods for constructing interval estimates have been slow to develop, as Eubank and Speckman (1993) noted. Recent papers (Fan *et al.* (1998) and Xia (1998)) have revisited this area suggesting improvements and new approaches.

Bowman and Azzalini (1997, pp.89-94) describe the construction of ‘reference

bands' for assessing the plausibility of *constant* and *no-effect* model fits to the data. These bands indicate where the nonparametric fit should lie when the simpler alternative is true. These methods capitalise on the unbiasedness of the local linear fit when the true function is linear and employ an estimate of the error variance, a topic we shall devote considerable attention to in Chapter 2. Bowman and Young (1996) describe related graphical comparisons of nonparametric curves in a number of contexts including repeated measures, survival analysis and binary response data. Similar bands are constructed in the context of nonparametric analysis of covariance (to be discussed in the next section) by Young and Bowman (1995).

### Model comparison tests

Although confidence bands aid in the assessment of the feature of data, these complement rather than replace formal tests of hypotheses regarding the form of the regression function. Of the tests which involve univariate smooths that have emerged, they tend to be motivated either by a comparison with a parametric regression fit to the same set of data or with another nonparametric regression fit. Although it is the latter context which is developed in the following Chapters, a brief overview of methods in both settings is given here, since many of the approaches from the former setting are applied (sometimes incorrectly) to the latter.

Using an analogy with linear regression, authors such as Cleveland and Devlin (1988) and Hastie and Tibshirani (1990, Section 3.9) have proposed the use of *approximate F tests* to compare different model fits to data. This is despite the fact that exact distribution results potentially do not hold when nonparametric fits are involved. They are, however, recommended as 'rough guides' to assessing the form of the regression function.

The approximate test statistic has the familiar form<sup>1</sup>:

$$F_{RSS} = \frac{(RSS_0 - RSS_1)/(df_0 - df_1)}{RSS_1/df_1},$$

---

<sup>1</sup>The notation is consistent with that used later in the thesis.

where  $RSS_0$  and  $RSS_1$  are the *residual sums of squares* of the null and alternate model fits being compared,  $df_0$  and  $df_1$  are *approximate error degrees of freedom* for the null and alternate model fits respectively. In the univariate setting this approach has been suggested to compare a smooth fit with parametric (constant and linear) fits. This test has also been proposed for comparing two versions of the nonparametric fit corresponding to different smoothing parameters, i.e. as a type of smoothing parameter selection method.

Recall that, in a fully parametric setting, this method is sometimes called the *Reduction method*. This reflects the requirement that the null and the alternate models are *nested*. That is, the null model must be a special case or ‘reduced form’ of the alternate model. This requirement is met when the null model is a parametric or a ‘smoother’ nonparametric version of the alternate nonparametric model. This concept is related to the definition of the *approximate degrees of freedom* which will be defined and discussed at length in Chapter 2.

Another requirement of this test is that the alternate model is *unbiased* and that the null model is also *unbiased* when the regression function is of the null model’s form. This condition holds for the local linear smoother when the null model is of a linear parametric form, but fails under a more general function (as shown in Section 1.2.1). These and other aspects of this test will be discussed at length in Chapter 3, so it is sufficient to note it here.

A more general approach to inference using univariate smooths is the *pseudo likelihood ratio test* (PLRT) introduced by Azzalini et al. (1989). As the name suggests, the test can be applied in the context of a range of parametric and nonparametric models, not just the normal-error univariate regression setting considered here. Azzalini and Bowman (1993) focus on this context, however, and demonstrate how the test can be used to compare a linear null model with a nonparametric alternate model. They propose the use of the *residuals* from the linear fit as the *responses* and (taking into account the correlation which is present amongst the residuals) proceed by testing for no-effect across the residuals.

The form of the test statistic is particularly straight forward, namely:

$$F_{PLRT} = \frac{RSS_0 - RSS_1}{RSS_1}.$$



The test rests on the fact that  $F_{PLRT}$  reduces to a quadratic form in zero mean normal variates when the null (linear/no-effect) model is true. Distributional properties of quadratic forms can then be used to determine the significance of an observed value.

Related proposals have been described by many authors, including: Raz (1990), Firth *et al.* (1991) and Härdle and Mammen (1993). Hart (1997) gives a book length treatment of the use of smoothing to aid in the assessment of the fit of *parametric* models.

Fewer authors have considered tests when both the null and the alternate models are nonparametric. Hall and Hart (1990) and King *et al.* (1991) are two of the earlier works which consider this. Young and Bowman (1995) made a significant contribution to this area when they considered nonparametric effects in the context of an analysis of covariance. Suppose there are  $p$  groups and within the  $i$ th group  $n_i$  pairs of a response variable and a single predictor variable (covariate) are observed. Of interest is whether the data observed in each group could have come from a common regression function. The test involves comparisons of the  $p$  smooths ( $\hat{m}_i(\cdot)$   $i = 1, \dots, p$  say) generated from the groups' data with the smoother derived from all the data (ignoring the grouping) yielding the estimate  $\hat{m}(\cdot)$ .

The test statistic proposed to compare these regression curve estimates is:

$$\frac{\sum_{i=1}^p \sum_{j=1}^{n_i} (\hat{m}_i(x_{ij}) - \hat{m}(x_{ij}))^2}{\hat{\sigma}^2}$$

where  $x_{ij}$  is the value of the  $j$ th covariate value in the  $i$ th group and  $\hat{\sigma}^2$  is an estimate of the error variance which is assumed to be constant across groups. When linear smoothers are used the test statistic can be shown to be a ratio in quadratic forms. Furthermore, the attractive design adaptive property of the local linear smooth ensures that the asymptotic biases in estimation of the  $p + 1$  curves are identical when the null hypothesis that regression function is the same across the groups, provided the same smoothing parameter is used for each fit.

With these properties noted, the test statistic reduces to a quadratic form in approximately standard normal variates. This is an approximation because although asymptotic bias cancels in the numerator, bias remains in the estimate

of  $\sigma^2$ . Noting the relatively small size of this bias component, Young and Bowman go on to describe the assessment of an observed value of the test statistic. Because of the similar forms of the test statistics, this assessment is identical to that used for the PLRT statistic described in Azzalini and Bowman (1993). For cases where there is evidence of differences between the groups' regression functions, Young and Bowman (1995) describe a modification of the test to compare groups' smooths for evidence that their regression functions are parallel.

This approach is not only of value in the analysis of covariance setting but suggests ways in which comparisons can be made amongst nonparametric models with more than one nonparametric term. These methods are developed, applied and assessed in Chapters 3, 4 and 5 in the context of nonparametric models which admit more than one predictor variable.

## 1.3 Multivariate Nonparametric Regression

### 1.3.1 Generalisations to a multidimensional setting

The material above has focused on the univariate predictor case. Rarely, in practice, though, is only one predictor variable available. This raises the question of smoothing techniques in multivariate settings. We introduce the notation  $(X1_i, X2_i, \dots, Xp_i, Y_i)$  for  $i = 1 \dots n$  to represent  $n$  realisations of a response variable  $Y$  and  $p$  predictor variables (covariates)  $X1, \dots, Xp$ . Let  $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p)$  be the set of observations, i.e.  $p$   $n$ -vectors containing the observed predictor values. There are several ways in which nonparametric models can relate the expectation of  $Y$  to more than one  $X$  covariate. This section describes some of these models and some of the issues surrounding them.

#### Multivariate smoothers

A natural extension is to assume that the response and predictors are related by:

$$E(Y|X1 = x1, X2 = x2, \dots, Xp = xp) = m(x1, x2, \dots, xp)$$

where  $m(\cdots)$  is a  $p$  dimensional smooth function. To estimate this ‘surface’ generalisations of many of the univariate smoothers exist, which will be referred to as *surface smoothers*. When applied to the responses these methods yield an estimate,  $\hat{m}(x_1, x_2, \dots, x_p)$ , of the regression surface.

The same principles of local smoothing described in the univariate setting can be applied in this setting although complications related to the increase in dimensions do arise. One such complication is the question of defining a local neighbourhood using  $p$  dimensional kernel functions. Consideration must be given to the possibility that the predictors are defined on different scales or that they are correlated. Approaches such as standardisation or defining local neighbourhoods using the predictor covariance matrix have been suggested to address these issues.

A more profound issue is the *curse of dimensionality* which states that as the dimensions of the design space increase neighbourhoods with a fixed number of points become less local (Bellman, 1961). One result of this ‘curse’ is that non-parametric surface estimators decrease in efficiency very quickly as the dimensions of the design space grow. To combat the curse requires that (linear) increases in dimensionality be accompanied by (exponential) increase in the data available in order to maintain the stability of the smooth surface estimates. This, clearly, is a real challenge and limits the application of such models to two or possibly three covariates at the most.

Another limitation of a multivariate smoothing approach is the difficulty in interpreting the estimate. Beyond two dimensions visualisation becomes increasingly limited in aiding the presentation and interpretation of the estimate. In addition, because no constraints are imposed on the form of the surface (apart from its smoothness) the potentially complex structure makes it difficult to separate out the effects of different covariates which is often the aim of multivariate modelling.

These factors give impetus to the search for techniques which incorporate the flexibility of smooths within a modelling framework which can summarise the main features and ideally facilitate inference. Such approaches are known as *dimensionality reduction techniques*, an important class of which is introduced in the following section.

## Additive models

One way around the curse of dimensionality is to restrict the form of the smooth. One approach which has gained in popularity over the last ten years is the restriction that the nonparametric effects of the predictors combine in an *additive* fashion rather than an arbitrary way. This gives rise to *additive models* which estimate regression functions by fitting models of the form:

$$\hat{m}(X1, \dots, Xp) = \hat{\alpha} + \hat{m}_1(X1) + \dots + \hat{m}_p(Xp). \quad (1.1)$$

where  $\hat{\alpha}$  is an estimate of the overall mean response and each of the  $m_j$  are univariate smooths on the respective predictor <sup>2</sup>.

This type of model avoids the difficulties associated with higher dimensions since it uses univariate smooths to estimate the components  $m_j(\cdot)$ , thereby avoiding the curse of dimensionality. The other attractive feature is its ease of interpretability. The effect of any single predictor, conditional on the other smooth functions in the model, is captured by a univariate smooth and can hence be plotted and examined with ease. Techniques for examining these ‘components’ of the additive model are the focus of the bulk of this thesis.

Of course, for the additive model fit to provide a meaningful description of the data, the additivity restraint must apply also to the true regression function, that is:

$$E(Y|X1, \dots, Xp) \equiv m(X1, X2, \dots, Xp) = \alpha + m_1(X1) + m_2(X2) + \dots + m_p(Xp)$$

This means that the model is nearly always an approximation to the true regression function. This has not been a barrier to its development or use however, since it has the potential to uncover important predictors and the way they influence the true function. Care must be taken when interpreting the fitted components however. For instance the additivity assumption must be checked and covariates examined for their effect on the response. Again, a large part of this thesis is

---

<sup>2</sup>The definition of additive models is, in fact, much broader than this, i.e. components are not required to be univariate or nonparametric, however this description will serve the purposes of introduction.

devoted to the description of techniques for these purposes.

### Other generalisations

A further generalisation of additive models is given by a class known as *Generalized Additive Models* (GAMs). GAMs are the nonparametric equivalents of *Generalized Linear Models* (GLMs, McCullagh and Nelder 1989). As such they allow for a variety of error distributions, not just the normal errors considered above. Indeed any distributions belonging to the exponential family are admissible. Furthermore, the relationship between the response and the predictors may be through a (known) link function. This introduces methods such as logistic and Poisson regression with smooth covariate effects.

However, a consequence of the greater versatility is that estimation requires iterative schemes. This introduces extra complexity into the analyses of the fitted values including methods of inference described above. As such, these models will not be considered in the current work. This is one instance where crude analogies with the parametric setting or bootstrap techniques offer the only inferential guidance available (see Hastie and Tibshirani (1990) for a comprehensive description).

Of course, this section has not attempted an exhaustive or comprehensive description of nonparametric models which tackle the ‘curse of dimensionality’. Other approaches include *projection pursuit regression* (Friedman and Stuetzle, 1981), *single index models* (Härdle *et al.*, 1993), *quasilikelihood semiparametric models* (Severini and Staniswalis, 1994) and *Generalized Partially Linear Single Index Models* (Carroll *et al.*, 1997). Each of these differ in the way they achieve dimensionality reduction, the flexibility of the regression function and the error distributions and links admissible. Reference will be made to these where necessary in the following chapters, although they are not the principal focus of the work.

### 1.3.2 Estimation

#### Multivariate smoothers

Local polynomials generalise easily, at least conceptually, to the task of estimating multivariate regression functions. For instance, a local linear approach in a bivariate setting proceeds by fitting *locally weighted planes* through the responses. Of course, this requires a bivariate nonnegative kernel function and specifications of the bandwidths in each direction to define the local weights. A common choice, and the one used throughout this thesis, is to define the kernel as the product of standard (univariate) normal kernels, one for each dimension, i.e. a bivariate normal density with  $\rho = 0$ , together with two bandwidths which scale the respective predictors. Although more general forms of the kernel function are possible, the simpler product kernel is generally sufficient and has the advantage that the resulting properties of the estimate resemble those in the univariate setting. A detailed description of the local linear bivariate smoother and its properties is given in Section 3.2.7.

Extensions of the univariate smoothing spline technique to two or more dimensions is less straight forward, however. Approaches such as the thin plate spline, discussed in detail in Green and Silverman (1994), and the multivariate tensor-product spline, studied by for example Friedman (1991), have been proposed. Given the nature of the models to be considered the generalisations offered by local polynomial models are far more appropriate and are used exclusively in the following chapters.

#### Additive models

As the form of Equation 1.1 implies, the method of estimating an additive model involves estimating several univariate smooth functions simultaneously. This, indeed, is the key to its ‘dimensionality reduction’ effect. This raises the issue of fitting additive models. Buja *et al.* (1989) describe an iterative technique called

*backfitting*<sup>3</sup>. To introduce this, consider the conditional expectation

$$E(Y - \alpha - \sum_{j \neq k} m_j(X_j) | X_k) = m_k(X_k)$$

which gives rise to the following algorithm<sup>4</sup>:

1. Initialise:  $\alpha = \bar{y}$ ,  $\hat{\mathbf{m}}_j = \hat{\mathbf{m}}_j^0, j = 1, \dots, p$
2. Cycle:  $j = 1, \dots, p, 1, \dots, p, \dots$   
 $\hat{\mathbf{m}}_j = \mathbf{S}_j(\mathbf{y} - \alpha - \sum_{k \neq j} \hat{\mathbf{m}}_k | \mathbf{x}_j)$
3. Continue (ii) until the fitted functions don't change.

This introduces the notation  $\mathbf{S}_j(\mathbf{y} | \mathbf{x}_j)$  to represent a univariate smoother applied to responses  $\mathbf{y}$  against the predictor  $\mathbf{x}_j$ , thus producing an estimate of  $m_j(\cdot)$  at the  $n$  observed values in  $\mathbf{x}_j$ , i.e.  $\hat{\mathbf{m}}_j$ . Since we shall always use linear smoothers, the operation can be expressed using a suitable  $n \times n$  smoothing matrix  $\mathbf{S}_j$ . At each step  $\hat{\mathbf{m}}_j$  is updated by removing the effects of all the other variables ( $k \neq j$ ) and then smoothing these *partial residuals* against  $\mathbf{x}_j$ . Since the smooth functions are each being estimated simultaneously, the iteration is necessary to achieve the correct partial residuals (and thus fitted components) eventually. A necessary constraint to ensure the identifiability and uniqueness of the final fit is that each smoother  $\mathbf{S}_j$  return fitted values with mean 0. Such smoothers are called *centred* and will be defined more precisely in Section 3.2.4.

As with the subjects of previous sections, there is much detail which could be described. Issues such as the set of equations which the backfitting algorithm is solving, the existence and uniqueness of a solution to these equations and the convergence properties of the algorithm have been the subject of much research (see Hastie and Tibshirani (1986), Buja *et al.* (1989)).

A special case, which is the focus of Chapters 3 to 5, is an additive model with two nonparametric components. It can be shown (Hastie and Tibshirani (1990)

<sup>3</sup>The idea of the algorithm was introduced in the context of *projection pursuit regression* by Friedman and Stuetzle (1981).

<sup>4</sup>Source: Hastie and Tibshirani (1990), Section 4.4

§5.3.4) that the estimating equations derived from the backfitting algorithm have explicit solutions in terms of the two component univariate smoothing matrices. This not only simplifies the conditions for the existence of solutions, but allows the properties of the fit to be investigated and these suggest ways in which inference using such a model may proceed.

Another special case of an additive model which is considered later is a semi-parametric model. These models contain a mixture of both nonparametric and linear components. Provided they do not have more than two covariates appearing in nonparametric terms, these too have explicit definitions of the fits which again pave the way to inference.

### 1.3.3 Smoothing parameter selection

Many of the ‘classical’ bandwidth selection techniques in the univariate setting, e.g. cross-validation and other techniques based on penalising ideas, can be extended to the multivariate setting. Herrmann *et al.* (1995) developed an iterative plug-in approach for use with bivariate kernel regression estimators. There has also been some recent activity in bandwidth selection methods for additive and semiparametric models with more than one nonparametric term. See, for instance, Opsomer and Ruppert (1999) and Simonoff and Tsai (1999).

Although methods are emerging in the multivariate context, these are not considered in detail in this current work for the same reasons as were given in Section 1.2.2 in the univariate setting. Instead, significance traces and size traces will be used to cover a range of smoothing parameters for the purposes of inference, thereby avoiding the delicate question of smoothing parameter selection which arises in the context of estimation.

### 1.3.4 Methods of inference

In a multivariate setting methods of inference are even more fundamental to their effective use. Because there is a number of potential covariates, a vital task is to identify those variables which contribute significantly to the observed responses and the form in which they combine. Taking the special case of two covariates (the focuses of Chapters 3 and 4) regression functions ranging from a constant



(no-effect in either covariate) model to a general bivariate smooth surface are possible. Therefore when we speak of model inference in the multivariate context we mean methods of determining the most appropriate form of the model to fit to the data.

The standard approach is to perform techniques analogous to those in the multivariate parametric setting, for instance techniques used in multiple regression and GLM settings. Additive models, with their obvious parallels with multiple regression, lend themselves particularly to this approach. Cleveland and Devlin (1988) pioneered the use of an approximation of the familiar F-test in the context of multidimensional additive models. The difficulty is that distributional results do not hold in the additive model context and therefore these ‘approximate’ methods (even with corrections) can only be used to provide vague guides to variable selection and other questions of inference (see Hastie and Tibshirani (1990) §6.8).

Bootstrap techniques have also been used to make comparisons amongst multivariate nonparametric models (Hastie and Tibshirani 1990, pp. 293). These approaches employ test statistics analogous to parametric settings and are usually in the ‘generalized’ setting where there is little or no distributional theory to work with. Once again, however, biased estimators will undermine the approaches and therefore we will focus attention on developing methods in simpler settings which attempt to address the bias problem directly.

Given the importance of this area, it is perhaps surprising that more attention has not focused on issues of inference. In order for nonparametric regression approaches to attain their full potential, inferential techniques must be developed alongside those of estimation. The major aim of this thesis is to develop and examine methods of inference via model comparisons in this context. The approach taken will be to adapt some of the methods in the univariate setting (Section 1.3.2) which compared different nonparametric fits to the same data for use in the multivariate setting.

## 1.4 Software

The previous sections have set the scene for the following chapters by introducing the concepts and methods of certain nonparametric models. This introduction, however, wouldn't be complete without a description of the software used throughout this work. This is indeed a fundamental aspect of nonparametric modelling in its own right, since the power and usefulness of these approaches have emerged with the rapid increase in the availability and speed of computing.

All of the code used throughout this thesis has been programmed in the S-Plus statistical software language. Although S-Plus possess nonparametric modelling capabilities, a library of S-Plus functions called 'sm', made available with Bowman and Azzalini (1997), has been used extensively. This was partly to gain a greater understanding and transparency in the computational methods used but also because S-Plus does not contain a local linear smoother with a fixed bandwidth (only a 'nearest neighbour' version). In addition to the 'sm' library, extensive use was made of the facilities in S-Plus to write specialised functions to implement the methods described herein. Indeed, behind every result presented in the following pages stands many tailored functions, the design and implementation of which are not acknowledged directly in the text. A summary and brief description of some of the main functions constructed are given in Appendix A.

## 1.5 Overview of Thesis

The chapters which follow divide into two sections. The first comprises Chapters 2 through 5 which explore several issues surrounding inference via nonparametric model comparisons. Chapter 2 starts by considering the estimation of the error variance in the nonparametric setting, considering particularly the case of two predictors. Chapter 3 extends methods of model comparisons beyond the univariate setting to the bivariate setting. Chapter 4 investigates the methods of Chapter 3 via a series of simulation studies. Chapter 5 then extends these results still further to the semiparametric setting with an unlimited number of linear terms and at most two terms appearing nonparametrically.

Chapters 6 and 7 are different in that they are motivated by applied problems.

Chapter 6 explores a set of environmental data. In the course of this investigation nonparametric models are used to explore covariate and seasonal effects. Chapter 7 describes a retrospective analysis of data from a medical context. Here again, nonparametric techniques are found to be useful in capturing features of data and form a key part of the analysis. Chapter 8 ends with a review and summary of the findings together with a discussion of the potential for further work.

## Chapter 2

# Error Variance Estimation in Bivariate Nonparametric Regression

### 2.1 The Rôle of the Error Variance

The last chapter gave an overview of the methods and issues related to estimation of an unknown regression function  $m(\cdot)$ . As Dette *et al.* (1998) put it, ‘without any doubt, the estimation of the regression function has been one of the most challenging fields during the past’. Thus, as the last chapter showed, the topic has received much attention in the research literature. However, for inference to proceed we clearly need more than just a point estimate, we also need a measure of the variability of the estimate.

Consider again the underlying univariate model which relates two variables  $X$  and  $Y$ ,  $n$  realisations of which are observed and stored in  $\mathbf{x}$  and  $\mathbf{y}$ :

$$\mathbf{y} = m(\mathbf{x}) + \boldsymbol{\varepsilon} \tag{2.1}$$

where  $\boldsymbol{\varepsilon}$  is an  $n$ -vector of zero mean random variables with constant variance  $\sigma^2$ . Recall that a linear smoother yields fitted values of the form  $\hat{\mathbf{m}} = \mathbf{S}\mathbf{y}$ , where  $\hat{\mathbf{m}}$  is an  $n$ -vector of fitted values and  $\mathbf{S}$  is an  $n \times n$  *smoothing matrix*. The  $n \times n$

variance-covariance matrix of the fitted values can be written as

$$\text{var}(\hat{\mathbf{m}}) = \text{var}(\mathbf{S}\mathbf{y}) = \mathbf{S}\text{var}(\mathbf{y})\mathbf{S}^T = \sigma^2\mathbf{S}\mathbf{S}^T,$$

where  $\text{var}(\mathbf{y})$  is  $\sigma^2\mathbf{I}_n$  ( $\mathbf{I}_n$  is the  $n \times n$  identity matrix).

This clearly illustrates the key rôle estimates of  $\sigma^2$  have in methods of inference. In fact, ‘...the estimation of the variance  $\sigma^2$  is nearly as important as the estimation of [the regression function] itself ...’ (Dette *et al.* (1998)), furthermore Buckley and Eagleson echoed similar sentiments back in 1988: ‘A great deal of effort has gone into developing estimators of the underlying regression  $m(\cdot)$  while the estimation of  $\sigma^2$  has been relatively ignored’. The estimation of  $\sigma^2$ , particularly in the bivariate case, is the focus of this chapter.

Before we consider this, however, one generalisation of the model described in Equation 2.1 should be noted. Heteroscedastic nonparametric models allow for the variance of the error terms to be a function of the independent variable(s). Ruppert *et al.* (1997) summarise the suggested approaches to variance-function estimation as:

1. Estimate the regression function
2. Initially, estimate the variability at design points  $X_i$  by the residual from the fit or alternatively define pseudoresiduals using a weighted average of a fixed number of the  $Y_i$ ’s (eg. difference-based methods)
3. Smooth the squared residuals or pseudoresiduals using kernel or local linear methods
4. Choose the bandwidths for smoothing the residuals or pseudoresiduals either subjectively or by a data based method.

They list Carroll (1982), Hall and Carroll (1989), Silverman (1985), Gasser *et al.* (1986) and Müller and Stadtmüller (1987) as the ‘sparse’ literature on this topic.

This topic is interesting in its own right and will surely develop into an important aspect of the approach to modelling. However, for the purposes of this thesis and specifically the issues of inference considered, it is sufficient to note this

approach and consider the homoscedastic model in detail.

## 2.2 Univariate Estimators

A useful starting point is to review existing approaches to error variance estimation in univariate nonparametric model settings. Two types of estimators have emerged in this setting. They both are functions of residuals but differ in how the residuals are obtained. The first class, RSS based estimators, define residuals in the familiar way as distances between a fitted regression curve and the observed responses. The second class defines the residuals by *differencing* the observed responses, i.e. no overall fit to the data is necessary. The term ‘pseudoresiduals’ was coined by Müller and Stadtmüller (1987) in the context of heteroscedastic models to describe the second type of residuals and has since been adopted (Ruppert *et al.* (1997)) and will be used in this thesis together with the term ‘difference based’ estimators to identify estimators based on them.

### 2.2.1 Residual sum of squares estimators

At the heart of the first class of variance estimators is the residual sum of squares (RSS) from a fitted nonparametric model, i.e.  $RSS = \mathbf{y}^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\mathbf{y}$ . By analogy with linear regression models, an obvious approach to the estimation of  $\sigma^2$  is to adjust RSS by a measure of the error degrees of freedom, that is:

$$\hat{\sigma}_{RSS}^2 = \frac{RSS}{df_{error}}. \quad (2.2)$$

This raises the question of what degrees of freedom correspond to a nonparametric fit to the data. The degrees of freedom of a fit reflect the degree of smoothing or, equivalently, the effective number of parameters used. This is an interesting quantity in its own right, as it defines a scale on which different smoothers can be calibrated with respect to the amount of smoothing they perform. In the context of variance estimation, however, its primary use is to define the *error degrees of freedom* of a fitted model. As in linear parametric models, this is defined to be the difference of the sample size  $n$  and the degrees of freedom of the smoother.

Hastie and Tibshirani (1990, pp 54) suggest  $df_{error} = n - \text{tr}(2\mathbf{S} - \mathbf{S}^T\mathbf{S})$  as the error degrees of freedom to be used in Equation 2.2. This expression comes from the expectation of the residual sum of squares, namely:

$$\begin{aligned}
 E(RSS) &= E(\mathbf{y}^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\mathbf{y}) \\
 &= E((\mathbf{m} + \boldsymbol{\varepsilon})^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})(\mathbf{m} + \boldsymbol{\varepsilon})) \\
 &= E(\mathbf{m}^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\mathbf{m} + \boldsymbol{\varepsilon}^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\boldsymbol{\varepsilon} + 2\mathbf{m}^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\boldsymbol{\varepsilon}) \\
 &= \mathbf{b}_S^T\mathbf{b}_S + E(\boldsymbol{\varepsilon}^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\boldsymbol{\varepsilon}) \\
 &= \mathbf{b}_S^T\mathbf{b}_S + E(\boldsymbol{\varepsilon}^T\mathbf{I}\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T\mathbf{S}^T\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T\mathbf{S}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T\mathbf{S}^T\mathbf{S}\boldsymbol{\varepsilon}) \\
 &= \mathbf{b}_S^T\mathbf{b}_S + \sigma^2n - \sigma^22\text{tr}\{\mathbf{S}\} + \sigma^2\text{tr}\{\mathbf{S}^T\mathbf{S}\} \\
 &= \mathbf{b}_S^T\mathbf{b}_S + \sigma^2(n - \text{tr}\{2\mathbf{S} - \mathbf{S}^T\mathbf{S}\})
 \end{aligned} \tag{2.3}$$

Clearly, if  $\mathbf{b}_S^T\mathbf{b}_S = \mathbf{m}^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\mathbf{m}$  (the sum of the squared biases over the design points) was negligible then  $\hat{\sigma}_{RSS}^2$  using  $df_{error}$  defined above would be a sensible estimator of  $\sigma^2$ . However, we have already noted that all nonparametric models contain bias and therefore  $\mathbf{b}_S^T\mathbf{b}_S > 0$ . This property that  $\hat{\sigma}_{RSS}^2$  overestimates  $\sigma^2$  in nonparametric regression is what motivates the work in the rest of this chapter.

It should be noted that two alternate definitions of degrees of freedom have been suggested (Hastie and Tibshirani, 1990 pp 52-55), namely  $\text{tr}\{\mathbf{S}\}$  and  $\text{tr}\{\mathbf{S}^T\mathbf{S}\}$ . The first is motivated by analogy with linear models and the relationship between the hat matrix and the number of parameters in the fitted model. If  $\mathbf{s}_i$  is the  $i$ th row of the smoothing matrix  $\mathbf{S}$  then it follows that the summed variances of the fitted values are given by:

$$\begin{aligned}
 \sum_i \text{var}(\hat{m}(x_i)) &= \sum_i \text{var}(\mathbf{s}_i\mathbf{y}) \\
 &= \sum_i \text{var}(\mathbf{s}_i\mathbf{m} + \mathbf{s}_i\boldsymbol{\varepsilon}) \\
 &= \sum_i \mathbf{s}_i\mathbf{s}_i^T\sigma^2
 \end{aligned}$$

$$\begin{aligned}
&= \text{tr}\{\mathbf{S}\mathbf{S}^T\}\sigma^2 \\
&= \text{tr}\{\mathbf{S}^T\mathbf{S}\}\sigma^2.
\end{aligned}$$

Although all three expressions are intuitive measures of the amount of smoothing performed, for the purposes of scaling  $RSS$  to define an estimator of  $\sigma^2$ ,  $n - \text{tr}(2\mathbf{S} - \mathbf{S}^T\mathbf{S})$  is clearly the most appropriate.

Estimators of this type employing kernel smoothers to obtain the fitted values have been suggested and investigated by a number of authors (see Hall and Carroll (1989) and Neumann (1994)). Cleveland and Devlin (1988) used this approach in the context of locally weighted regression, drawing largely on the analogy with the linear parametric model. Hall and Marron (1990) define an estimator based on  $RSS$  which they show to be asymptotically first and second order optimal. Authors such as Wahba (1978, 1983) and Carter and Eagleson (1992) derived conceptually similar estimators using a roughness penalty (smoothing spline) approach to model fitting.

The performance of these estimators in the univariate case has been well researched, and they have been found to possess some attractive properties. One complication to their use is their direct dependence on the fit to the data which introduces an unavoidable degree of bias into the estimator. Another concern is the choice of smoothing parameter used to estimate the regression function. This introduces a level of arbitrariness: what is an appropriate choice of the smoothing parameter for the purposes of estimating  $\sigma^2$ ?

Despite these drawbacks, this approach has also received widespread acceptance, perhaps because of its familiarity from the linear model setting. In their comprehensive text, Hastie and Tibshirani (1990) only define estimators of  $\sigma^2$  in terms of the residual sum of squares. For this reason, we shall consider  $RSS$  based estimators of  $\sigma^2$  throughout this thesis.

### 2.2.2 Difference based estimators

There exists a class of estimators of  $\sigma^2$  which do not require the regression function to be estimated explicitly. These estimators work by removing the trend in the



data by *differencing* the response values via various schemes, and hence are known as *difference based estimators*. Different forms of these exist in the univariate case, three of which are briefly described here: namely those of Rice (1984), Gasser *et al.* (1986) and Hall *et al.* (1990).

Rice (1984) considered *first order* differences via the estimator

$$\hat{\sigma}_R^2 = \frac{1}{2(n-1)} \sum_{i=2}^n (Y_{[i]} - Y_{[i-1]})^2$$

( $Y_{[i]}$  is the response associated with the  $i$ th *largest* predictor value  $X_{[i]}$ ). Gasser *et al.* (1986) suggested a *second order* estimator

$$\hat{\sigma}_G^2 = \frac{2}{3(n-2)} \sum_{i=3}^n \left( \frac{1}{2} Y_{[i-2]} - Y_{[i-1]} + \frac{1}{2} Y_{[i]} \right)^2.$$

Hall *et al.* (1990) introduced an asymptotically optimal difference-based estimator of variance using a *difference sequence*  $\{d_i\}_{i=-m_1, \dots, m_2}$  with

$$\sum d_j = 0$$

$$\sum d_j^2 = 1$$

where  $d_{-m_1}, d_{m_2} \neq 0$ ,  $m_1, m_2$  are non-negative integers and  $r = m_1 + m_2$  denotes the *order* of the variance estimator

$$\hat{\sigma}_{H,r}^2 = \frac{1}{n-r} \sum_{k=m_1+1}^{n-m_2} \left( \sum_{j=-m_1}^{m_2} d_j Y_{[j+k]} \right)^2.$$

Each of these estimators can be written as  $\mathbf{y}^T \mathbf{D}^T \mathbf{D} \mathbf{y} / \text{tr}\{\mathbf{D}^T \mathbf{D}\}$  where  $\mathbf{D}$  is an  $n \times n$  matrix which depends only on the design space. For Rice's and Gasser's estimators, respectively, these are:

$$\mathbf{D}_R = \frac{1}{\sqrt{2(n-1)}} \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 & 0 \\ & & & \cdots & & & \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}$$

and

$$\mathbf{D}_G = \frac{2}{\sqrt{3(n-2)}} \begin{bmatrix} 0 & & \cdots & & & & & 0 \\ 0 & & \cdots & & & & & 0 \\ -1/2 & -1 & 1/2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1/2 & -1 & 1/2 & 0 & \cdots & 0 & 0 \\ & & & \cdots & & & & \\ 0 & & \cdots & & 0 & -1/2 & -1 & 1/2 \end{bmatrix}$$

When applied to the response vector  $\mathbf{y}_{\square}$ , the *ordered* (by  $X$ ) vector of responses, then  $\mathbf{D}$  yields a vector of *pseudoresiduals*,  $\mathbf{D}\mathbf{y}_{\square}$ , and hence the terminology. The advantage of these approaches is that they do not require an explicit fit, and no smoothing parameter is needed<sup>1</sup> nor a specific type of smoother. Furthermore, they often have a small bias for small sample sizes (Dette *et al.* (1998), pp 754-755) and are computationally very simple.

However, these methods aren't without their challenges. These estimators do not achieve the asymptotic optimal rate of some 'residual based' estimators (see Eagleson (1989), Hall and Marron (1990)), although their performance can be comparable. For this reason, some authors have focused attention on their finite sample properties (Seifert *et al.* (1993)). Furthermore, the choice of the method and order of differencing which minimises the (finite sample) MSE of the estimator is not straightforward.

Ruppert *et al.* (1997) argue that since pseudoresiduals are based on a fixed

---

<sup>1</sup>Dette *et al.* (1998) show that the order of the differences has an influence which is comparable with the smoothing parameter

(independent of  $n$ ) number of design points they are correlated, even asymptotically, which complicates their analysis. This is true in the univariate setting but, as we shall see in Section 2.5, this is not true in the bivariate setting. We shall see that the pseudoresiduals are not defined using a fixed number of points, and therefore this argument against pseudoresiduals does not apply in this context. Furthermore, their focus was on heteroscedastic models, involving smooths of the pseudoresiduals, unlike here where we seek to estimate a constant variance.

Within the class of difference based estimators, finite sample and asymptotic properties have to be balanced. The estimators of Gasser *et al.* (1986) and Hall *et al.* (1990) have come to represent these two ‘approaches’ (finite sample and asymptotic respectively, e.g. Seifert *et al.* (1993), Dette *et al.* (1998)). Through a detailed simulation study, Dette *et al.* (1999) developed guidelines for the choice of an estimator of  $\sigma^2$ . They considered three classes of estimators which corresponded to the difference based estimators of Gasser *et al.* (1986), those of Hall *et al.* (1990) and the kernel type (RSS based) estimator of Hall and Marron (1990). Although factors such as the sample size, the magnitude of the residual variance and the order of differencing employed influence the performance of the different estimators, the regression function itself was found to be the most influential property. The Hall *et al.* (1990) estimator was found to perform poorly with regression functions with high oscillations, and thus in the absence of information concerning the magnitude of the oscillations of the regression function, or faced with extremely noisy data, the estimator of Gasser *et al.* (1986) is preferable.

## 2.3 Variance Estimation in Higher Dimensions

The discussion of variance estimators in the previous section focused on the problem in the univariate case. Extensions to the bivariate and higher-dimension settings have been considered by several authors, although these are confined to special cases and have not received the depth of treatment which the univariate estimators have received.

### 2.3.1 RSS based approaches

Several authors have noted that techniques based on residual sums of squares in the univariate setting may be applied in higher dimensions. Cleveland and Devlin (1988) apply their RSS based approach in the context of one, two and three independent variables. Buckley *et al.* (1988) comment ‘our theorems on the form of the optimal estimator [of  $\sigma^2$ ] apply equally to any ... multivariate nonparametric regression where a quadratic roughness penalty is used’. Likewise, Ruppert *et al.* (1997) state that ‘in principle’ their theory on a general class of variance estimators based on local polynomial smoothing can be extended to multivariate predictor variables.

In contrast to the majority of treatments which deal with the higher dimensional setting with a passing comment, however, an early work by Breiman & Meisel (1976) considers estimating ‘intrinsic variability’ solely in a multivariate setting. Their approach, also based on RSS, is a somewhat primitive version of local linear modelling whereby the design space is progressively divided into subspaces. This continues until there is evidence that a linear model applied to each subregion is satisfactory. Their estimator of  $\sigma^2$  is then the sum of the RSS’s corresponding to the linear fits within the subregions divided by the number of design points.

As was the case in one dimension, these estimators all require an estimate of the regression function, and thus are explicitly linked to the choice of smoothing parameter(s). It was noted in the one dimensional case that the choice of the smoothing parameter presents a challenge to their routine use. In higher dimensions this problem is more acute, since there are a number of dimensions each of which require some statement of the level of smoothing.

A related problem in the multi-dimensional setting is the *curse of dimensionality* which states that as the dimensions increase the amount of data required to obtain valid estimates of the regression surface increases exponentially. The impact on smoothing parameter selection is to force parameters of sufficient size to guarantee that a sensible number of responses is used to estimate the surface at each point. This has a knock-on effect in that the RSS estimator based on such a fit could potentially contain considerable bias.

This may suggest that *variable bandwidth* regression estimators, such as Cleveland's (1979) LOWESS and Cleveland & Devlin's (1988) loess estimators, may be of particular use in helping regulate the degree of smoothing throughout the design space. These procedures work by taking a fixed number of the design points around each evaluation point. That is, the smoothing parameter is the *span* of the kernel function in terms of the proportion of data covered, rather than a fixed bandwidth as is often used in local polynomial and kernel methods. These approaches are sometimes referred to as  $k$ -nearest-neighbour methods where  $k$  is the fixed number of design points used to calculate the regression estimator at each evaluation point.

RSS estimators based on smooth fits achieved via this approach do suffer a major drawback, which perhaps explains why they are not considered in the literature on univariate RSS based estimators. As Gasser & Seifert (1994) comment, there is a lack of mathematical underpinning of these methods. They also state that 'a design-adaptive bandwidth implies a more complex and potentially larger bias, compared to a fixed bandwidth'. We have stressed the importance of minimising bias in the fit to achieve an accurate estimator of  $\sigma^2$  and therefore it is natural to avoid an approach which potentially may yield more biased estimators.

A third reason for not considering RSS estimators using these smoothers, is the context of model inference in which we seek to use the estimators of  $\sigma^2$ . As we shall see in Chapter 3, a fixed bandwidth local linear approach to model fitting yields attractive asymptotic bias results which motivate the approaches to model comparisons considered later. As such, we will confine ourselves to RSS based estimators of  $\sigma^2$  using fits of the same form.

### 2.3.2 Difference based approaches

Approaches using pseudoresiduals, defined by differencing responses, have the appeal that they do not require an explicit estimate of the regression surface, and therefore do not require the specification of a smoothing parameter. Furthermore, they incorporate the 'nearest-neighbour' idea and therefore adjust to the scarcity of the points in a region. Extensions to the higher dimensional settings are not straightforward though. Several approaches in the bivariate setting will be

described in the remainder of this section and developed further in the next.

A difference based approach in the bivariate setting is discussed by Hall *et al.* (1991) where they estimate the residual variance in the context of two-dimensional signal processing. They propose an estimator along the lines of that presented in Hall *et al.* (1990) extended to two dimensions. The authors note, 'a major dissimilarity between the one- and two-dimensional cases is the very rich variety of different configurations [of design points] available in two dimensions' (Hall *et al.* (1991), pp 477). As such, the paper focuses on the selection of optimal configurations, and of optimal difference weights for given configurations. Because of its context, the major restriction of the approach is the requirement that the design points form a lattice. Within the class of lattice designs, however, the results are in fact very general, holding without change for different lattice types. Extensions to higher-dimension lattices are also possible. The limitation to lattice designs, however, is not realistic in the context we seek to use them, i.e. inference using nonparametric model fits.

Seifert *et al.* (1993), at the end of their treatment of univariate estimators, comment that 'difference-based methods can easily be generalised to multidimensional designs.' They, too, highlight the very rich class of configurations and also the increasing portion of the boundary for growing dimension as issues particular to the higher dimension setting. An example of extending the univariate approach advocated in Gasser *et al.* (1986) to the bivariate case, was sketched by Herrmann *et al.* (1995). In a univariate setting, the variance estimator is a function of pseudoresiduals defined as the distance from a response  $y_i$  to the line connecting the responses at the design points either side of the  $x_i$ . In the bivariate case, pseudoresiduals are defined as the mean difference between  $y_i$  and planes through configurations of three 'neighbouring' design points and the variance estimator is defined as a weighted average of the squared pseudoresiduals.

This raises the important question of the idea of 'nearest neighbours' in higher dimensions. In one dimension it is obvious which two points to interpolate to define a pseudoresidual. In two dimensions, however, it is not so straight forward since there may be more than one triplet of design points which could be used to define a pseudoresidual at each design point. The questions of how to identify suitable triplets and then how to derive a single pseudoresidual from all these

points, therefore, arise in the bivariate context.

Herrmann *et al.* (1995) define an estimator for the special case of a rectangular grid design, where the symmetry and orthogonality makes it clear which way to proceed. They indicate that their approach applies to non-equidistant/non-orthogonal designs but do not describe this in any detail. Clearly, in a non-equidistant/orthogonal design space expressions for the pseudoresiduals are design dependent and thus ‘general’ results do not exist. A literature search has failed to reveal any treatment of non-equidistant/orthogonal designs in connection with the estimation of  $\sigma^2$ . Herrmann *et al.* (1995) offer some suggestions on the way to proceed in these settings, which we shall return to in Section 2.5 where a difference based estimator in the bivariate setting is defined and studied.

Finally, in an as yet unpublished work, Dette *et al.* (1999) extend the treatment of variance estimators to the multivariate setting. Their approach is to consider the response as multivariate ( $d$  dimensional, say) regressed on one predictor. Thus, the aim is the estimation of a  $d \times d$  variance-covariance matrix  $\Sigma$ , rather than the scalar  $\sigma^2$ . Our focus is on the reverse, i.e. one response variable with multiple (especially two) predictors. Therefore, unless the results of Dette *et al.* (1999) can be adapted, they are of little relevance to the focus of this chapter.

## 2.4 Improved RSS Based Estimator of $\sigma^2$

Equation 2.3 showed that the inherent bias in nonparametric fits affects the accuracy of the RSS based estimator of  $\sigma^2$ . This highlights how necessary it is to minimise the bias of the fitted model in order to obtain accurate estimates of  $\sigma^2$ . In this Section we shall investigate several adjustments to  $\sigma_{RSS}^2$  which attempt to reduce the influence of the bias term. Although they are not restricted to the bivariate case, they will be investigated in this context via a simulation study in Section 2.7.

### 2.4.1 Undersmoothing

Chapter 1 showed how the asymptotic bias of a univariate local linear smoother increases as the smoothing parameter increases. In Chapter 3, where a bivariate

class of models is considered in detail, it will be shown that this property holds for the bivariate local linear smoother as well. This suggests that an improved estimator of  $\sigma^2$  can be obtained by using  $RSS$  from a model fit with a *smaller* value of the smoothing parameter than that appropriate for estimating the underlying regression function. We shall refer to this modification as *undersmoothing* since a smaller smoothing parameter yields a more variable (i.e. less smooth) estimate of the regression function.

This approach highlights an aspect of smoothing parameter selection which has been noted by several authors. For instance Bowman and Azzalini (1997 pp 93) note that the importance of the choice of smoothing parameter varies depending on whether the smooth is to be used for the purposes of estimation or inference. In estimating the regression function, the trade off between bias and variance drives the choice of the smoothing parameter whereas for inference this trade off is less crucial. The consequence is that a range of smoothing parameter values will often return consistent results when used for the purpose of inference via the comparison of the fitted models, as the size traces presented in Chapter 4 will show.

The question of what degree of reduction to employ and its effect on the accuracy on the estimator of  $\sigma^2$  remain. It was noted in Chapter 1 that as the smoothing parameter decreases the local linear fit to the data approaches an interpolation of the observations. There therefore exists a lower bound of the smoothing parameter, beyond which the  $RSS$  becomes zero. Nevertheless, the use of residuals obtained by *undersmoothing* the data to estimate the error variance has been suggested by several authors (e.g. Hastie and Tibshirani (1990) pp. 48). The results from a simulation study, presented in Section 2.7, attempt to illustrate and quantify the improvements achieved by decreasing the smoothing parameter.

### 2.4.2 Double smoothing

Recall from Chapter 1 that the asymptotic bias of a local linear fit is a function of the second derivative of the (unknown) regression function. This feature suggests the use of a smooth of the *residuals* as the source of an estimate of  $\sigma$ . If most



of the curvature of the true underlying model is reflected in the fitted model, the residuals from this fit  $((\mathbf{I} - \mathbf{S})\mathbf{y})$  should be relatively free from dramatic curvature. In other words, the residuals themselves can be thought of as random fluctuations around a surface of less curvature than the original surface, but with approximately the same variability.

Consider applying the smoother to the residuals, i.e.  $\mathbf{S}(\mathbf{I} - \mathbf{S})\mathbf{y}$ , (i.e. *double smoothing* the original data). Provided less bias is contained in the smooth of the residuals than in the smooth of the original data, the residuals from the second smooth potentially have better properties to estimate  $\sigma^2$ . To see this recall that the bias term in the standard RSS is  $\mathbf{b}^T\mathbf{b} = ((\mathbf{I} - \mathbf{S})\mathbf{m})^T(\mathbf{I} - \mathbf{S})\mathbf{m}$  whereas, using the double smooth approach, the corresponding term is  $\mathbf{b}^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\mathbf{b}$  which it is reasonable to expect will be less than the first term.

This idea leads to the following definitions.

$$\begin{aligned}\hat{\sigma}_{DS}^2 &= \frac{\mathbf{y}^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})\mathbf{y}}{df_{DS}} \\ &= \frac{\mathbf{y}^T\mathbf{B}_{DS}\mathbf{y}}{df_{DS}}\end{aligned}\tag{2.4}$$

Methods using residuals from an initial smooth, i.e.  $(\mathbf{I} - \mathbf{S})\mathbf{y}$ , appear in other areas of the smoothing literature. One instance is where a semiparametric fit to the data is required. Speckman (1988) derived an estimator of the parametric part which uses partial residuals obtained via smoothing both the responses and the linear covariates and using these to estimate the linear component. Its advantage in that setting was to improve the asymptotic convergence and the asymptotic bias of the estimator of the parametric component. This is discussed at length in Chapter 5.

Of course, the previous action of using a smaller smoothing parameter can also be applied in conjunction with the double smoothing idea. These approaches are investigated in Section 2.7.

## 2.5 Defining a Difference Based Estimator in Two Dimensions

In this section we shall consider the difference based approach to estimating the error variance in a bivariate setting. Extensions of the univariate estimators of Gasser *et al.* (1986) sketched by Herrmann *et al.* (1995) are used as a starting point, developed and investigated in the context of design points which are a realisation of a random configuration of points in the  $(X, Z)$  plane. Special cases, such as the rectangular grid design, will be presented as illustrations.

The first step is to define pseudoresiduals at each design point. At a particular point,  $(X_i, Z_i)$  say, the pseudoresidual is defined as follows:

1. determine the set of  $n_i$  *neighbouring triplets* of points
2. for each triplet, calculate the distance between the plane which interpolates the triplet's responses and  $Y_i$ , thereby defining  $n_i$  *triplet distances*  $(\tilde{\epsilon}_{i1}, \dots, \tilde{\epsilon}_{in_i})$
3. define the pseudoresidual at the design point to be the average of these triplet distances, i.e.  $\tilde{\epsilon}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \tilde{\epsilon}_{ij}$

Neighbouring triplets are defined in terms of the *Delaunay triangulation* of the design space (Ripley 1981). Delaunay triangulations are created by joining design points who share an edge defined by Dirichlet tessellations. These tessellations partition the entire design space and consist of polygons constructed around each point to define regions that are closer to that point than any other. Thus, by joining points with neighbouring Dirichlet cells, *neighbouring points* are identified in a sensible way.

Having defined the neighbouring points for a given design point  $(X_i, Z_i)$ , neighbouring *triplets* of design points are three neighbours such that there is a path of two Delaunay edges connecting the three design points forming a triplet. In other words, the points in the triplet must be 'neighbours' themselves. One other requirement of the triplet points is that they don't lie on a straight line, for obvious reasons of the identifiability of the plane through the triplet responses.

In summary, each triplet of points associated with design point  $(X_i, Z_i)$  satisfies the following criteria:

- ◊ each point is connected with  $(X_i, Z_i)$  by a single edge (as defined by the Delaunay triangulations)
- ◊ there is a path of two edges (of the Delaunay triangulation) connecting the three design points forming a triplet
- ◊ the three design points are not located on a straight line.

As an illustration, Figure 2.1 shows the four triplets (four ‘corners’) corresponding to the interior point of a rectangular design<sup>2</sup>. The pseudoresidual at this interior point, defined above as the average of the distances from the response to the four planes, has an explicit formula given by

$$\bar{\varepsilon}_{i,j} = Y_{i,j} - \frac{1}{2}(Y_{i-1,j} + Y_{i+1,j} + Y_{i,j-1} + Y_{i,j+1}) + \frac{1}{4}(Y_{i-1,j-1} + Y_{i-1,j+1} + Y_{i+1,j-1} + Y_{i+1,j+1}).$$

where  $Y_{i,j}$  is the response at the design point defined by coordinates  $(i, j)$  in Figure 2.1.

It can be shown that different expressions apply at the boundary points of the rectangular design grid. For example, at the  $(i = 1, j = 1)$  corner point, the pseudoresidual is defined as

$$\bar{\varepsilon}_{1,1} = Y_{1,1} - (Y_{1,2} + Y_{2,1} - Y_{2,2})$$

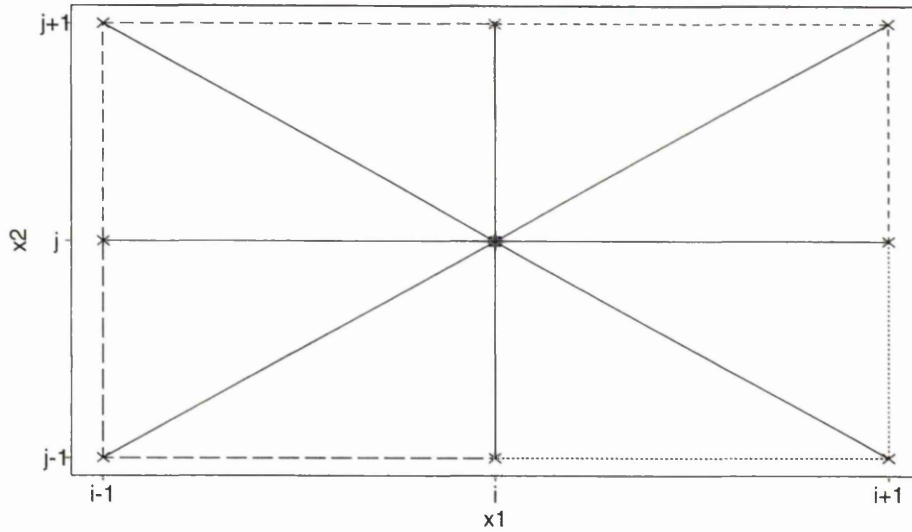
and at a point  $(i = 1, j)$  on the edge the pseudoresidual is

$$\bar{\varepsilon}_{1,j} = Y_{1,j} - Y_{2,j} - \frac{1}{2}(Y_{1,j+1} + Y_{1,j-1} - Y_{2,j+1} - Y_{2,j-1}).$$

Similar results can be derived for other corner and boundary points. Note that these expressions do not involve the spacing between points and thus hold for any rectangular grid, regular or irregular, regardless of spacing.

---

<sup>2</sup>Technically, the Delaunay triangulation is not defined for a rectangular grid, but the configurations shown here are natural ones and in the spirit of the Delaunay triangulation idea



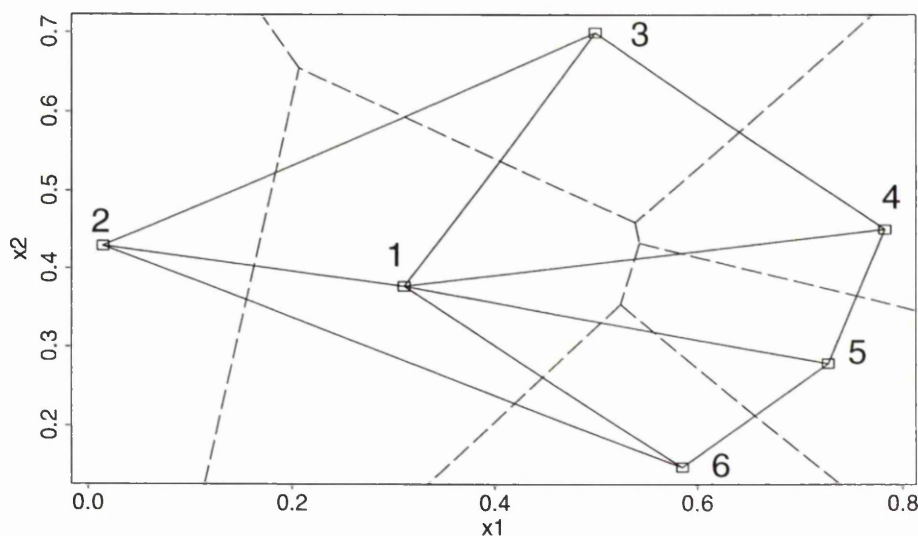
**Figure 2.1.** Representation of a segment of a rectangular grid showing an interior point ( $\square$ ) with 8 neighbouring points ( $\times$ ). The four triplets, used to define a pseudoresidual at the interior point, are shown using different line styles.

Figure 2.2 illustrates the definition of the neighbouring triplets with a realisation of a random design. The broken lines define the Dirichlet cells and the solid lines define the Delaunay triangulation. For the realisation shown, the pseudoresidual at design point 1 is defined using the set of triplets: (2,3,4), (3,4,5), (4,5,6), (5,6,2), (6,2,3), namely the average distance between the response at '1' and the planes interpolating the triplets' responses.

Having defined the pseudoresiduals for each point as the average distance between the response and the 'triplet planes', the error variance estimator is defined as a linear combination of squared pseudoresiduals. Herrmann *et al.* (1995) suggest a straight average:

$$\hat{\sigma}_{DIFF}^2 = \frac{1}{n} \sum_{i=1}^n a_i \tilde{\varepsilon}_i^2, \quad (2.5)$$

where  $a_i = \frac{\sigma^2}{\text{var}(\tilde{\varepsilon}_i)}$ . The reasoning of such an estimator can be seen by noting that



**Figure 2.2.** Realisation of a random design showing the Delaunay triangulation which define neighbouring triplets.

$$\begin{aligned} E(\hat{\sigma}_{DIFF}^2) &= \frac{1}{n} \sum_{i=1}^n \sigma^2 \frac{E(\tilde{\varepsilon}_i^2)}{\text{var}(\tilde{\varepsilon}_i)}, \\ &= \sigma^2 \end{aligned}$$

if  $E(\tilde{\varepsilon}_i) = 0$ , since then  $\text{var}(\tilde{\varepsilon}_i) = E(\tilde{\varepsilon}_i^2)$ . That is, for the difference based estimator to be unbiased the pseudoresiduals need to have expectations of zero.

Recall that  $\tilde{\varepsilon}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \tilde{\varepsilon}_{ij}$ , the average of the distances between the  $i$ th response and the planes through responses at neighbouring triplets. Clearly, the positioning of these ‘neighbouring responses’ with respect to the  $i$ th response, will determine properties such as the expectation of  $\tilde{\varepsilon}_i$ . According to the nonparametric model, the neighbouring responses are random deviations from some underlying regression surface and therefore it is the local curvature of this unknown surface which will determine the accuracy of the estimator.

If the regression surface in the region defined by the neighbouring triplets around the  $i$ th point can be approximated linearly, then the expectation of  $\tilde{\varepsilon}_i$

will be approximately zero. Conversely, if there is substantial local curvature in the regression surface (relative to the spacing between neighbouring points) then the interpolated planes are crude estimates of the regression function and there is potential for considerable bias to be present in the estimator. This estimator can be expected to perform best, then, when the underlying function is approximately linear in the space defined by the neighbouring triplets.

Hence, there are similarities with the performance of the RSS based estimator using a local linear smoother. The second derivative of the regression function in the asymptotic bias of the local linear fit show that it too is approximately unbiased if the regression function is locally linear. This suggests that the behaviour of the two estimators may therefore be similar. The simulations in Section 2.7 will explore this.

Returning to the definition of the estimator, however, to calculate  $a_i$  we note that  $\tilde{\varepsilon}_i$  is an average of linear combinations of the  $Y_i$ 's. Since the model states that  $\text{var}(Y_i) = \sigma^2$ , calculation of  $a_i$  requires a record of the constants used to define  $\tilde{\varepsilon}_i$  in terms of the  $Y_i$ 's. For instance, from the definitions of the pseudoresiduals defined above for the rectangular grid design it follows that  $a_i = 4/9$  for interior points,  $a_i = 1/4$  for corner points and  $a_i = 1/3$  for edge points.

It is possible that design points in certain realisations of random designs may not have three points satisfying the above criteria. As such, they do not yield pseudoresiduals and thus  $a_i$  is undefined. These points are analogous to the two end points  $X_{[1]}$  and  $X_{[n]}$  in the univariate case. However, the number of such points is not fixed and may in fact be zero. At these points the pseudoresidual and the  $a_i$  are set to zero and the  $n$  in Equation 2.5 is replaced with the number of points at which pseudoresiduals exist,  $n_{\tilde{\varepsilon}}$  say.

Since the pseudoresiduals are linear combinations of the responses, there exists a  $n_{\tilde{\varepsilon}} \times n$  matrix,  $\mathbf{D}$  say, such that the  $n_{\tilde{\varepsilon}} \times 1$  vector of pseudoresiduals is given by

$$\tilde{\varepsilon} = \mathbf{D}\mathbf{y}.$$

Let  $\mathcal{A}_{n_{\tilde{\varepsilon}}}$  be the diagonal matrix with  $a_i/n_{\tilde{\varepsilon}}$  as the  $i$ th diagonal element, that is

$\mathcal{A}_{n_\varepsilon} = \frac{1}{n_\varepsilon} \frac{1}{\text{diag}(\mathbf{D}\mathbf{D}^T)}$ , then

$$\hat{\sigma}_{DIFF}^2 = \mathbf{y}^T \mathbf{D}^T \mathcal{A}_{n_\varepsilon} \mathbf{D} \mathbf{y}.$$

It is significant that this estimator can be written as a quadratic form since, as we shall see in Chapter 3, this assists in the derivation of reference distributions when  $\hat{\sigma}_{DIFF}^2$  is used to define a test statistic.

Let us consider one modification of  $\hat{\sigma}_{DIFF}^2$  which attempts to improve its performance. Recall that at the  $i$ th design point there are  $n_i$  triplet distances which combine to define the pseudoresidual  $\tilde{\varepsilon}_i$ . Herrmann *et al.* (1995) suggest that an (unweighted) average of these distances be used. This implies that each triplet distance is equally valuable in providing information about the true unknown residual at this point. This is clearly not so. Consider, once again, Figure 2.2 where only one of the neighbouring triplets *surround* design point 1, namely triplet (6,2,3). The plane through the responses of triplet (6,2,3) is therefore more likely to approximate the regression function over point 1 than, say, the plane through the responses at (4,5,6), since clearly an extrapolation beyond the domain defined by the triple is required. Other factors such as the distances from the triplets to the design points and even the areas enclosed by the triplets will be related to the amount of information contained in each triplet distance.

For this reason, we propose defining the pseudoresiduals as weighted averages of triplet distances, that is,

$$\tilde{\varepsilon}_i = \frac{\sum_{j=1}^{n_i} w_{ij} \tilde{\varepsilon}_{ij}}{\sum_{j=1}^{n_i} w_{ij}},$$

where the  $w_{ij}$  reflect the precision of information contained in  $\tilde{\varepsilon}_{ij}$ . Although there are many weighting schemes which could be derived, a natural one, in the spirit of the weighted average of squared pseudoresiduals which defines the estimator, is to use

$$w_{ij} = \sigma^2 / \text{var}(\tilde{\varepsilon}_{ij}).$$

This weighting scheme has the effect of down-weighting those triplet distances which contain a high amount of variability and vice versa.

To appreciate the nature of such a weighting scheme in terms of the observed data, recall how the  $\tilde{\epsilon}_{ij}$  are defined. Let  $(X_{ij1}, Z_{ij1})$ ,  $(X_{ij2}, Z_{ij2})$ ,  $(X_{ij3}, Z_{ij3})$  be the design points comprising the  $j$ th triplet around the  $i$ th design point. Let  $A_{ij1}$  be the first row of

$$A_{ij} = \begin{bmatrix} 1 & X_{ij1} - X_i & Z_{ij1} - Z_i \\ 1 & X_{ij2} - X_i & Z_{ij2} - Z_i \\ 1 & X_{ij3} - X_i & Z_{ij3} - Z_i \end{bmatrix}^{-1}$$

then

$$\tilde{\epsilon}_{ij} = y_i - A_{ij1} \cdot \begin{bmatrix} y_{ij1} \\ y_{ij2} \\ y_{ij3} \end{bmatrix} = W_{ij}^T \begin{bmatrix} y_i \\ y_{ij1} \\ y_{ij2} \\ y_{ij3} \end{bmatrix}$$

where  $W_{ij}^T = [1 | -A_{ij1}]$ . Given this notation, the definition of weights follows directly as

$$w_{ij} = \sigma^2 / \text{var}(\tilde{\epsilon}_{ij}) = (W_{ij}^T W_{ij})^{-1}.$$

Figure 2.3 shows the weights of each triplet around point 1 of Figure 2.2. The weights are shown alongside the *middle* point of each triplet. They show that weights proportionate to the variance of the triplet distances yield consistent values in the sense that triplets which are *closer* to point 1 and which *surround* point 1 receive the greater weighting.

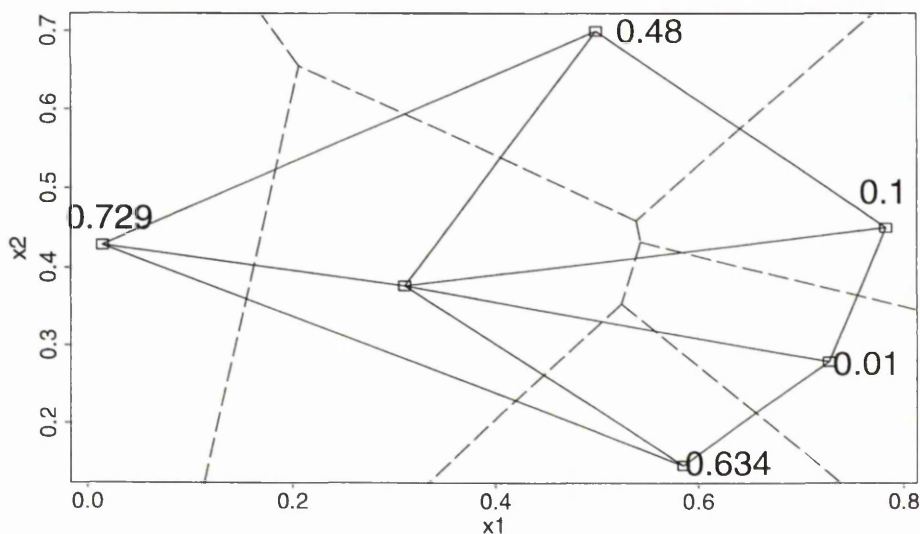
In matrix notation, this definition of the pseudoresiduals leads to

$$\tilde{\epsilon} = \mathbf{D}_{WT} \mathbf{Y},$$

where matrix  $\mathbf{D}_{WT}$  has the same dimensions as  $\mathbf{D}$  above and whose constants reflect the weighting scheme. These pseudoresiduals, squared and averaged (weighted), yield the estimator

$$\hat{\sigma}_{WT}^2 = Y^T \mathbf{D}_{WT}^T \frac{1}{n_{\tilde{\epsilon}}} \frac{1}{\text{diag}(\mathbf{D}_{WT} \mathbf{D}_{WT}^T)} \mathbf{D}_{WT} Y.$$





**Figure 2.3.** Realisation of design points shown in Figure 2.2 with the weights applied to the five triplet distances used to define the pseudoresidual at the centre point (weight is shown alongside the middle point of each triplet).

Hence the two estimators,  $\hat{\sigma}_{DIFF}^2$  and  $\hat{\sigma}_{WT}^2$  differ in the way they define the pseudoresiduals. Both can be expressed in the quadratic form  $\mathbf{y}^T \mathbf{A} \mathbf{y}$ , however, which allows their respective properties to be derived and compared. This is the topic of the next section.

## 2.6 Properties of Estimators

Having defined a number of estimators of the error variance in the bivariate setting, attention now turns to investigating their properties. Initially we shall focus on exact results which explicitly define the bias and variance of the estimators. Secondly, the results of a simulation study will be used to further highlight properties of the estimators.

### 2.6.1 Finite sample bias and variance

Each of the estimators introduced is of the form  $\hat{\sigma}^2 = \mathbf{y}^T \mathbf{A} \mathbf{y}$  where  $\mathbf{A}$  is an  $n \times n$  *estimating matrix*. This common form is particularly useful since it allows comparisons of estimators to reduce to comparisons of their respective estimating matrices. Furthermore, the distributional properties of a *quadratic form* random variable are accessible (see Chapter 29 of Johnson and Kotz 1972).

In the context of univariate modelling, Seifert *et al.* (1993) showed that an estimator of this form has the property,

$$E(\hat{\sigma}^2) = \sigma^2 \text{tr}(\mathbf{A}) + \mathbf{m}^T \mathbf{A} \mathbf{m}.$$

If the regression function is known and the design points are fixed, this expression allows for the direct comparison of different estimators based on finite samples without simulations. When interest lies in *random designs* the design dependent  $\mathbf{A}$ 's can be calculated over a number of simulated realisations of the design and their corresponding expectations averaged. We shall employ these results in Section 2.7 where the performances of a number of estimators under different conditions are investigated through a simulation study. The expectations of the estimators are presented together with the estimates from each simulated data set.

### 2.6.2 Asymptotic properties

Attempts have been made to consider the asymptotic properties of difference based estimators in the bivariate setting. Most notably, Hall *et al.* (1991) discuss a bivariate estimator analogous to an estimator they propose for the univariate setting (see Hall *et al.* (1990)) described briefly in Section 2.2.2. The definition of the estimator is driven by finding configurations of points and difference weights which minimise the asymptotic MSE of the estimator.

At first glance, this suggests that a consideration of the asymptotic properties may be useful in deciding between estimators. However, the expressions for asymptotic MSE employed by Hall *et al.* (1991) are based on the result that squared bias is asymptotically negligible relative to error about the mean, i.e.

$(E(\hat{\sigma}^2) - \sigma^2)^2 / \text{var}(\hat{\sigma}^2) \rightarrow 0$  as  $n \rightarrow \infty$ . They use the signal processing context in which they propose this estimator to support this assumption. They stress that

‘since images are typically large, with many finely spaced pixels, then the asymptotic theory [of negligible bias] is certainly relevant in practical problems where noise variance is estimated from parts of the image which are relatively smooth’.

This highlights two assumptions which cannot be guaranteed to hold in the model inference context in which we seek to use these estimators. Certainly we cannot assume that the number of design points observed will be sufficiently large or finely separated to allow for the asymptotic result to hold. Furthermore, although we too assume a degree of smoothness in the underlying function, this is not guaranteed over random designs which we wish to consider which differ from the regular grids of pixels available in signal processing.

We therefore note the existence of asymptotic results in the special case of finely spaced grids of design points, but conclude that the small sample properties of the previous section are more relevant to the context considered here.

## 2.7 Simulation Study

The previous sections have described a number of estimators of  $\sigma^2$ , including their small sample properties. This section will investigate the performance of these estimators under varying conditions via a simulation study, the results of which were summarised in Bock (1999).

### 2.7.1 Estimators under investigation

Five estimators of  $\sigma^2$  will be considered:

1.  $\hat{\sigma}_{RSS}^2$ : RSS based estimator using a model fit of the same form as the true regression function
2.  $\hat{\sigma}_{RSS2D}^2$ : RSS based estimator using a model fit of the most general form of the regression function (two dimensional smooth surface)

3.  $\hat{\sigma}_{DS}^2$ : RSS based estimator using double smoothing and a model fit of the same form as the true function
4.  $\hat{\sigma}_{DIFF}^2$ : difference based estimator
5.  $\hat{\sigma}_{WT}^2$ : difference based estimator using weighted triplets to define pseudo-residuals

Each of the RSS based estimators are calculated using three levels of the smoothing parameters which highlight the dependency between the performance of the estimator and the smoothing parameter used and therefore the effect of undersmoothing. The RSS type estimators using a fitted model of the same form as the *true* regression function can be thought of as ‘best case scenarios’ in the sense of their being parsimonious fits to the data. In the methods of inference via model comparisons discussed in Chapter 3, however, the RSS from a more general model fit are often used. Therefore, we include  $\hat{\sigma}_{RSS2D}^2$  to assess the effect of using the most general model fit in this context, namely bivariate smoothing, on the RSS based estimator of  $\sigma^2$ .

In total, therefore, 11 estimates of the error variance will be calculated from each data set: 3×3 RSS based and 2 difference based. Data sets will be simulated over a range of conditions which will be used again in Chapter 4 so the next section will describe them in some detail. Section 2.7.3 summarises the results of simulation study.

## 2.7.2 Simulated conditions

Each estimator will be applied to data generated by adding random  $N(0, \sigma^2)$  noise to a regression function evaluated at  $n$  design points from a particular design space. The following conditions will be considered:

◇ different regression functions

1. univariate model:  $m_u(X_1, X_2) = m_1(X_1)$
2. semiparametric model:  $m_s(X_1, X_2) = m_1(X_1) + X_2$
3. additive model:  $m_a(X_1, X_2) = m_1(X_1) + m_2(X_2)$

4. bivariate smooth model:  $m_b(X_1, X_2) = -X_1 + \Psi(X_1, X_2, 0.3, 0.3, 0.1, 0.1, 0) + X_2 - \Psi(X_1, X_2, 0.7, 0.7, 0.1, 0.1, 0)$

where,  $m_1(X_1) = 1 + X_1 - 0.75 \exp(-0.5(X_1 - 0.5)^2/0.01)$ ,  $m_2(X_2) = 1 + X_2 + 0.75 \exp(-0.5(X_2 - 0.5)^2/0.01)$  and  $\Psi(X_1, X_2, \mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$  denotes the probability density function of a bivariate normal distribution. Each regression function is scaled such that the range of values over  $[0, 1]^2$  is 1 before random noise is added.

◇ different design spaces

1. regular square grid over  $[0, 1]^2$
2. random (uniform) over  $[0, 1]^2$
3. random (bivariate normal,  $\mu_1 = \mu_2 = 0.5$ ,  $\sigma_1 = \sigma_2 = 0.15$ ,  $\rho = 0$ )
4. random (bivariate normal,  $\mu_1 = \mu_2 = 0.5$ ,  $\sigma_1 = \sigma_2 = 0.15$ ,  $\rho = 0.5$ )

◇ different sample sizes

1. 49 (suits a  $7 \times 7$  grid)<sup>3</sup>
2. 100 (suits a  $10 \times 10$  grid)

◇ different error variance,  $\sigma^2$

1. 0.01
2. 0.04

The choice of sample size and error variance represent low and moderate values encountered in practice. The regression functions were chosen such that they each exhibited departures from linear models. They are shown in Figure 2.4 evaluated over the regular grid.

---

<sup>3</sup>Only for regular grid

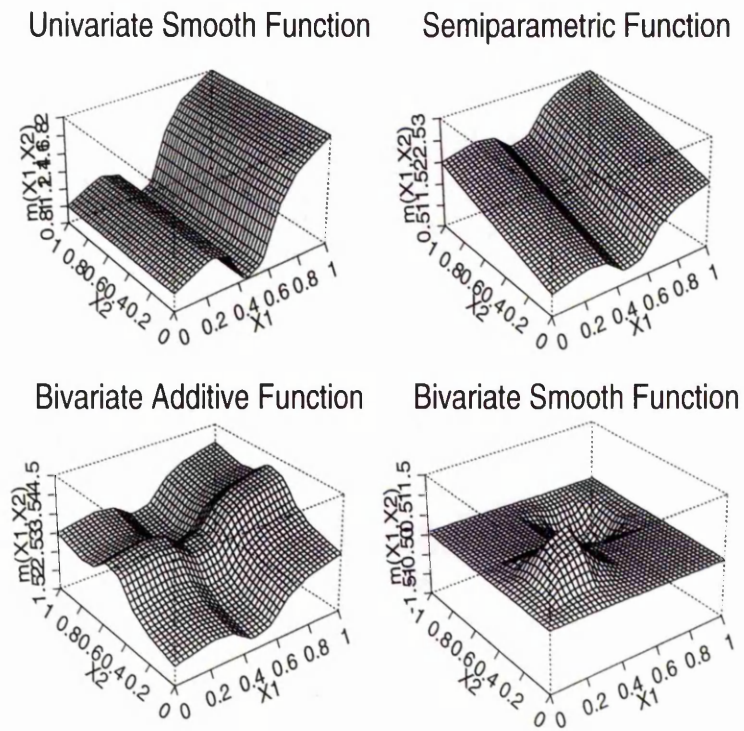


Figure 2.4. True underlying functions used to simulate data.

### 2.7.3 Simulation results

Under each of the settings described above, data sets were simulated and estimates of  $\sigma^2$  obtained using the 11 estimators listed in Section 2.7.1. The results from the simulations are presented in two forms:

- ◊ boxplots showing the distribution of the estimated values of  $\sigma^2$
- ◊ graphical summaries of the true bias of the estimators, calculated using the true regression surface

#### Regular grids

In a regularly spaced grid all the triplets surrounding a point are the same ‘distance’ from that point. Therefore the ‘weighted-triplets’ difference based estimator will be the same as the standard differenced based estimator and thus is omitted from the results for the regular designs. Also, when the underlying regression function is a general two dimensional surface, the ‘true’ and ‘general’ forms of the model are the same and thus the results are only presented once.

Figures 2.5 and 2.6 show the distribution of estimates over a regular grid for the two sample sizes (49 and 100) respectively. Each boxplot represents 500 simulations. Together with the bias properties of the estimators, summarised in Figure 2.7, these results exhibit the following properties:

**differenced based estimators** exhibits favourable bias properties. In terms of bias, it is comparable with the best RSS based estimator. The price it pays for its accuracy is its variance which is amongst the highest of the estimators considered.

**standard RSS based estimates** show clearly the effect of the choice of smoothing parameter on the performance of the estimator. As the smoothing parameter increases, the bias in the estimates increases markedly. A comparison of  $\hat{\sigma}_{RSS}^2$  and  $\hat{\sigma}_{RSS2D}^2$  shows the effect of misspecifying the form of the underlying function. Even for small smoothing parameters,  $\hat{\sigma}_{RSS2D}^2$  can contain a noticeable degree of bias.

**double smoothed estimates** offer some improvement over the standard RSS estimator. The degree of improvement increases as the smoothing parameter increases, thus acting as a buffer against increasing bias.

### Random designs

Figures 2.8 - 2.10 show the distribution of estimates over three random designs using a sample size of 100 points. Figure 2.8 is based on uniform-random design configuration whereas Figures 2.9 - 2.10 simulates over bivariate normal designs with correlation coefficients 0 and 0.5 respectively. In each case 500 simulations are represented.

The bias properties of the estimators, conditional on the realisations of the designs, are summarised in Figure 2.11. Since there was considerable agreement between the biases of the standard and bivariate smooth RSS estimators, only the former are presented.

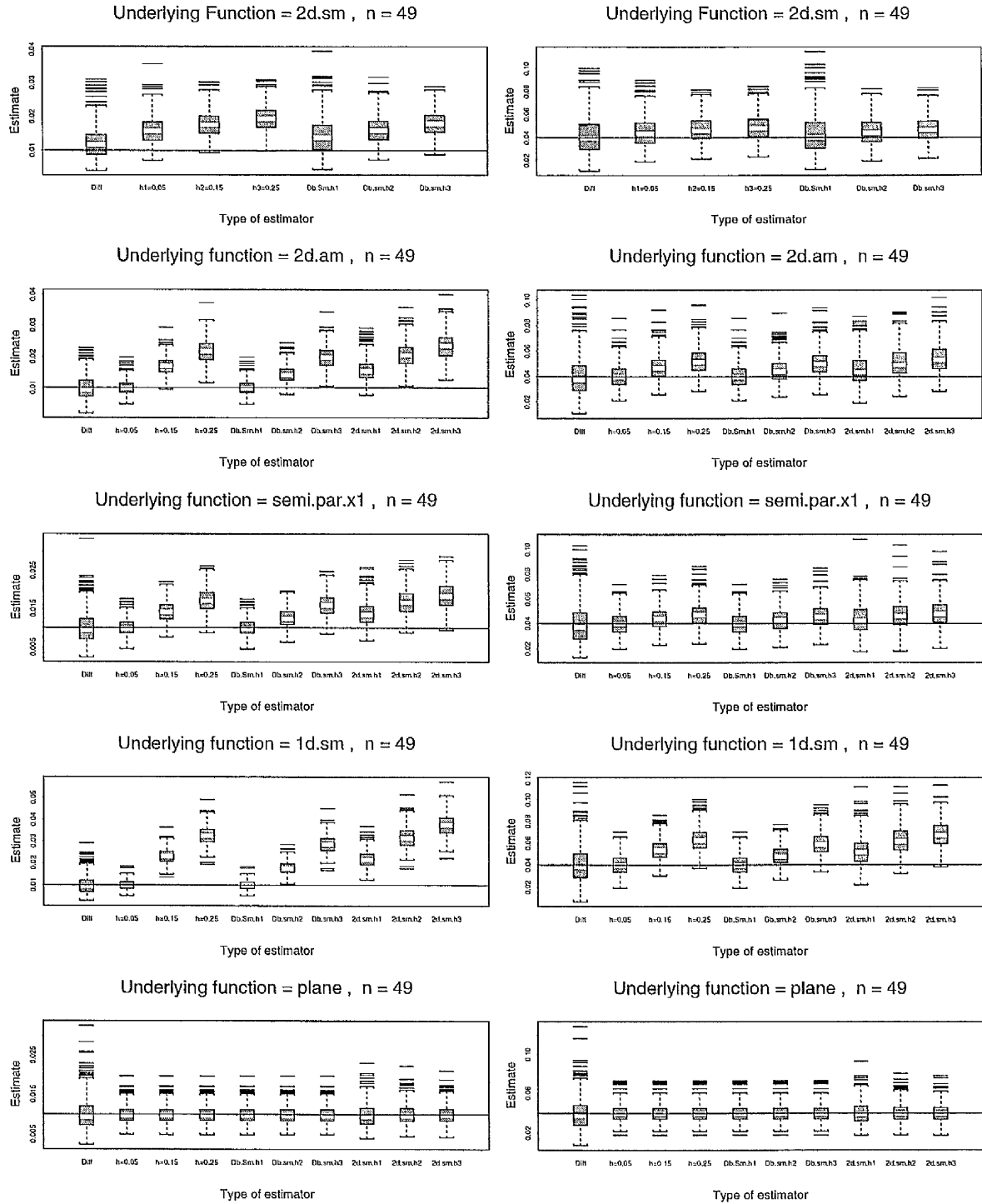
In summary, the results over the random designs exhibit the following properties:

**differenced based estimates** exhibit favourable bias properties. This is true for both the weighted and unweighted triplet versions, where there is little to separate them in terms of their bias performance. They are both comparable with the best RSS based estimator. One noticeable consequence of the random designs is that the variance of the difference based estimators are now comparable to the double smoothed RSS estimators.

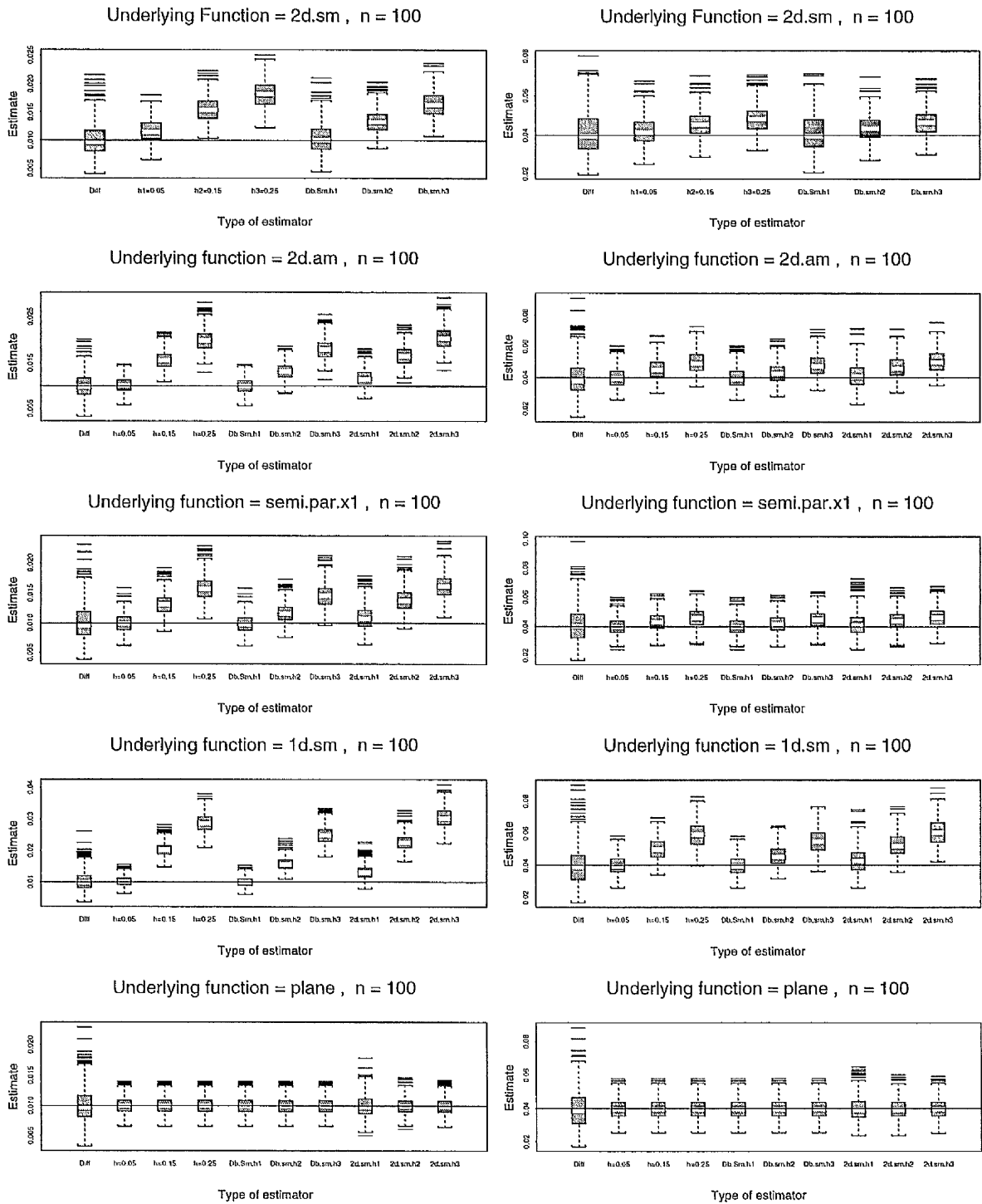
**standard RSS based estimates** show clearly the effect of the choice of smoothing parameter on the performance of the estimator. As the smoothing parameter increases, the bias in the estimates increases markedly. A comparison of  $\hat{\sigma}_{RSS}^2$  and  $\hat{\sigma}_{RSS2D}^2$  results shows that there is little ill-effect due to using a model fit of a more general underlying function. That is, the random distribution of design points results in the good performance of the bivariate smooth fits for the purposes of estimating  $\sigma^2$ .

**double smoothed estimates** show that these offer some improvement over the standard RSS estimator.

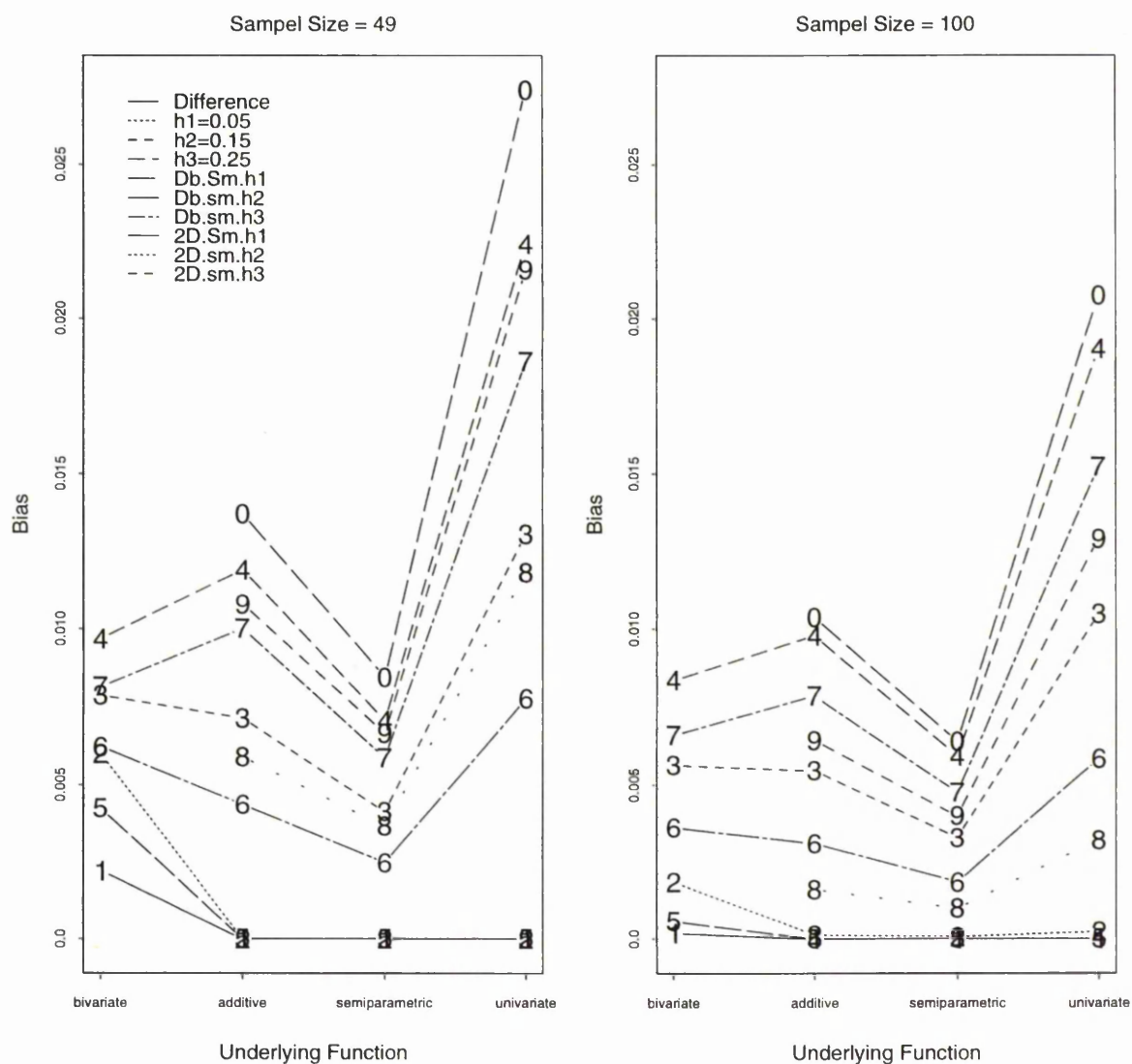




**Figure 2.5.** Boxplots showing the distribution of estimates of  $\sigma^2$  (solid horizontal line) from different estimators. 500 simulations of data generated over a regular grid were used.



**Figure 2.6.** Boxplots showing the distribution of estimates of  $\sigma^2$  from different estimators under different simulation conditions. 500 simulations of data generated over a regular grid were used in each boxplot.



**Figure 2.7.** Plots showing the finite sample biases of different estimators of  $\sigma^2$  under different regression functions and sample sizes over a regular grid.

**form of the random designs** does not appear to affect the performance of the estimators. Similar results are observed for each of the random designs considered.

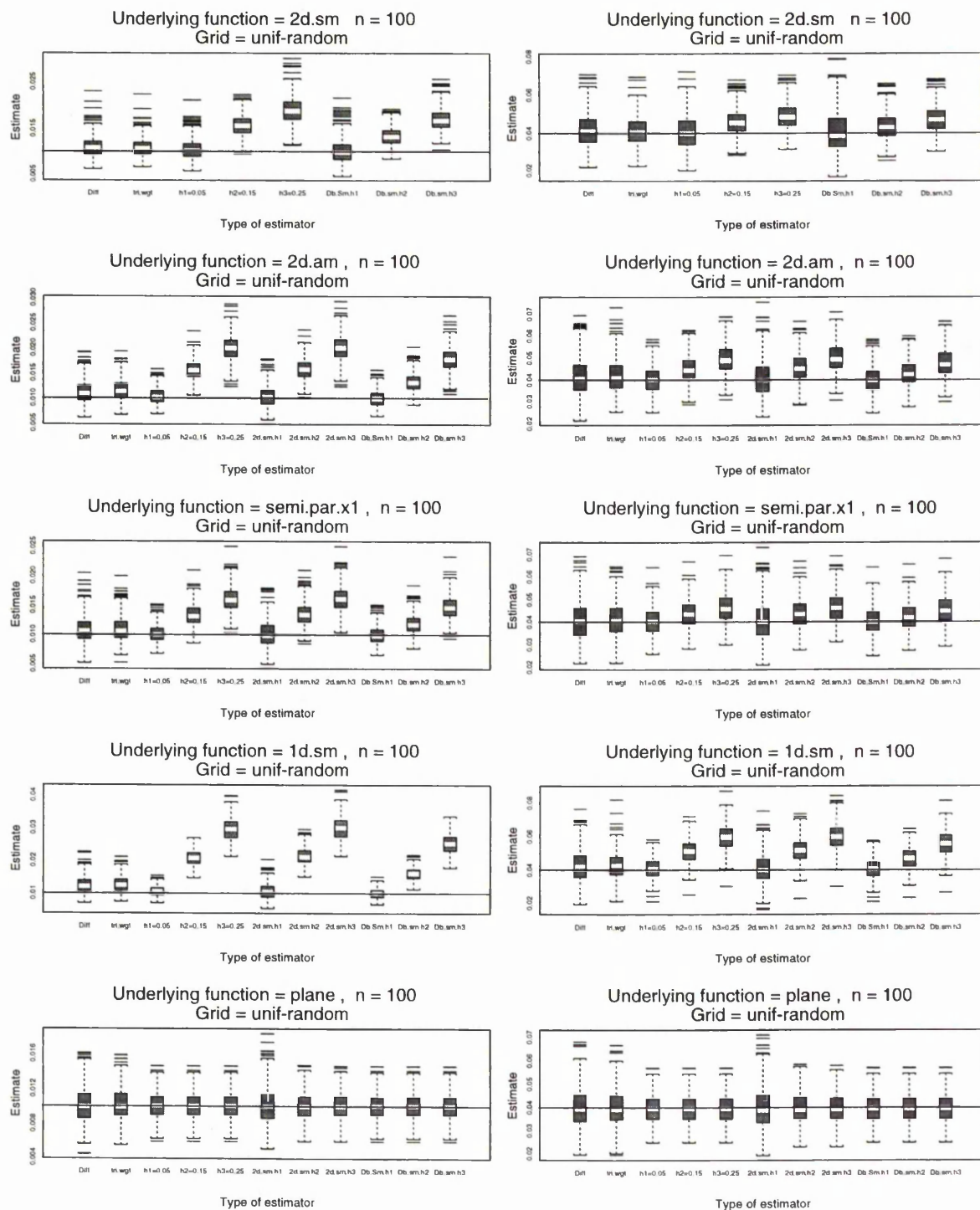
## 2.8 Discussion

This chapter has focused on the task of estimating  $\sigma^2$  in a bivariate regression setting. It was motivated by the challenges of inference using nonparametric model fits, particularly in the two dimensional setting, which will be considered in detail in the following three chapters. Several approaches to the estimation of  $\sigma^2$  have been discussed at length. Standard *residual based* estimators have been described and potential improvements investigated. As an alternative, *difference based* estimators have been developed and their performances investigated.

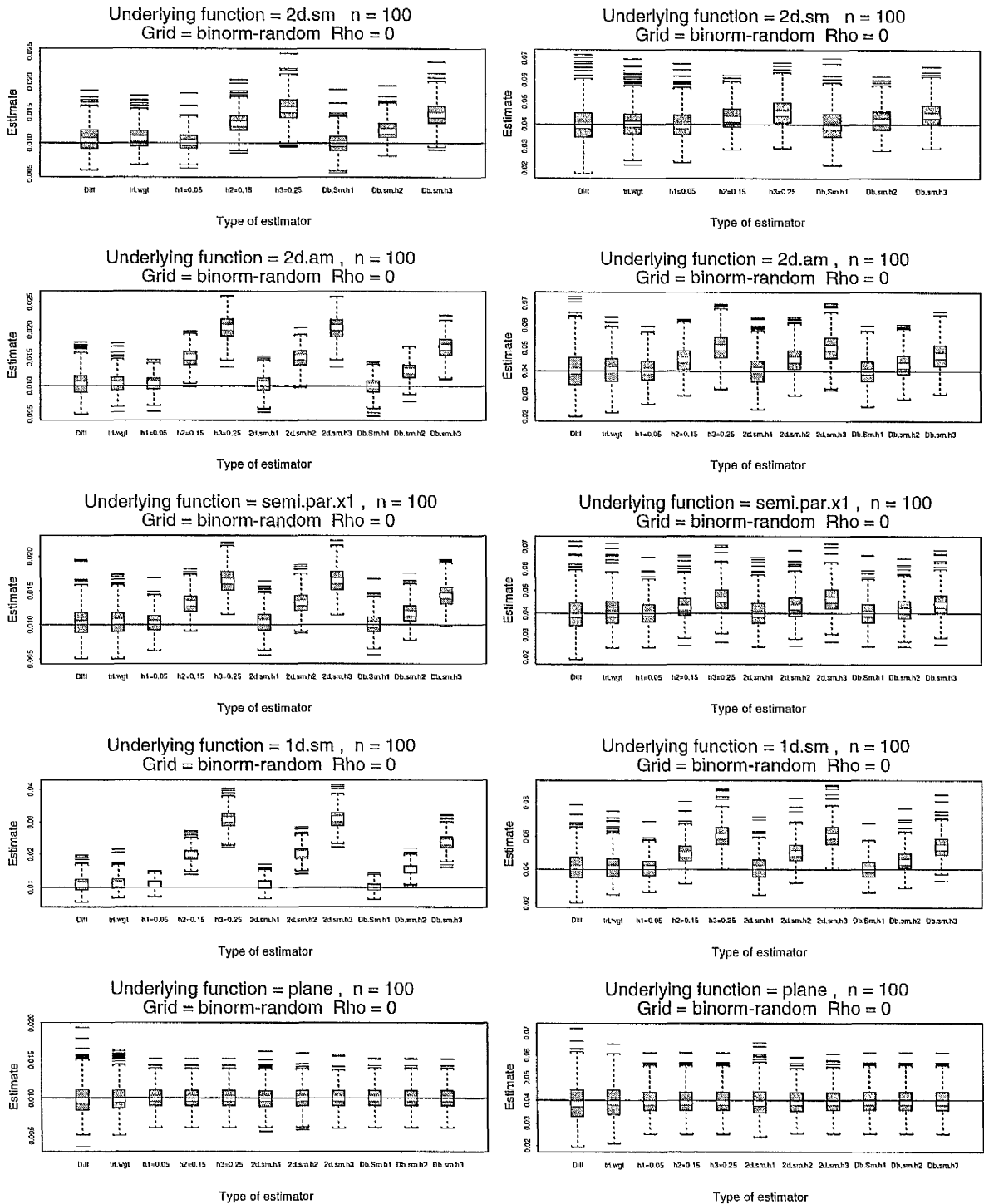
The results have clearly illustrated the rôle that the choice of smoothing parameter plays in the performance of the residual based estimators. Although improvements in accuracy are possible through the use of techniques such as double smoothing, the most critical factor remains the choice of smoothing parameter. This was highlighted in the results for the undersmoothed RSS based estimator which showed clearly that the finite sample biases were proportional to the smoothing parameter(s) used to define the explicit model fit.

When using RSS based estimators of  $\sigma^2$  the question of which smoothing parameter to use always arises. Even the observation that undersmoothing yields an improved  $\hat{\sigma}^2$  (in terms of bias) carries with it the open question of what degree of undersmoothing to use. Here lies an advantage of a difference based estimator, i.e. this question doesn't arise. In terms of accuracy relative to the best residual based estimators, difference based estimators were demonstrated to be comparable over random designs and slightly less efficient over regular grids, although in some settings a high price was paid in terms of precision.

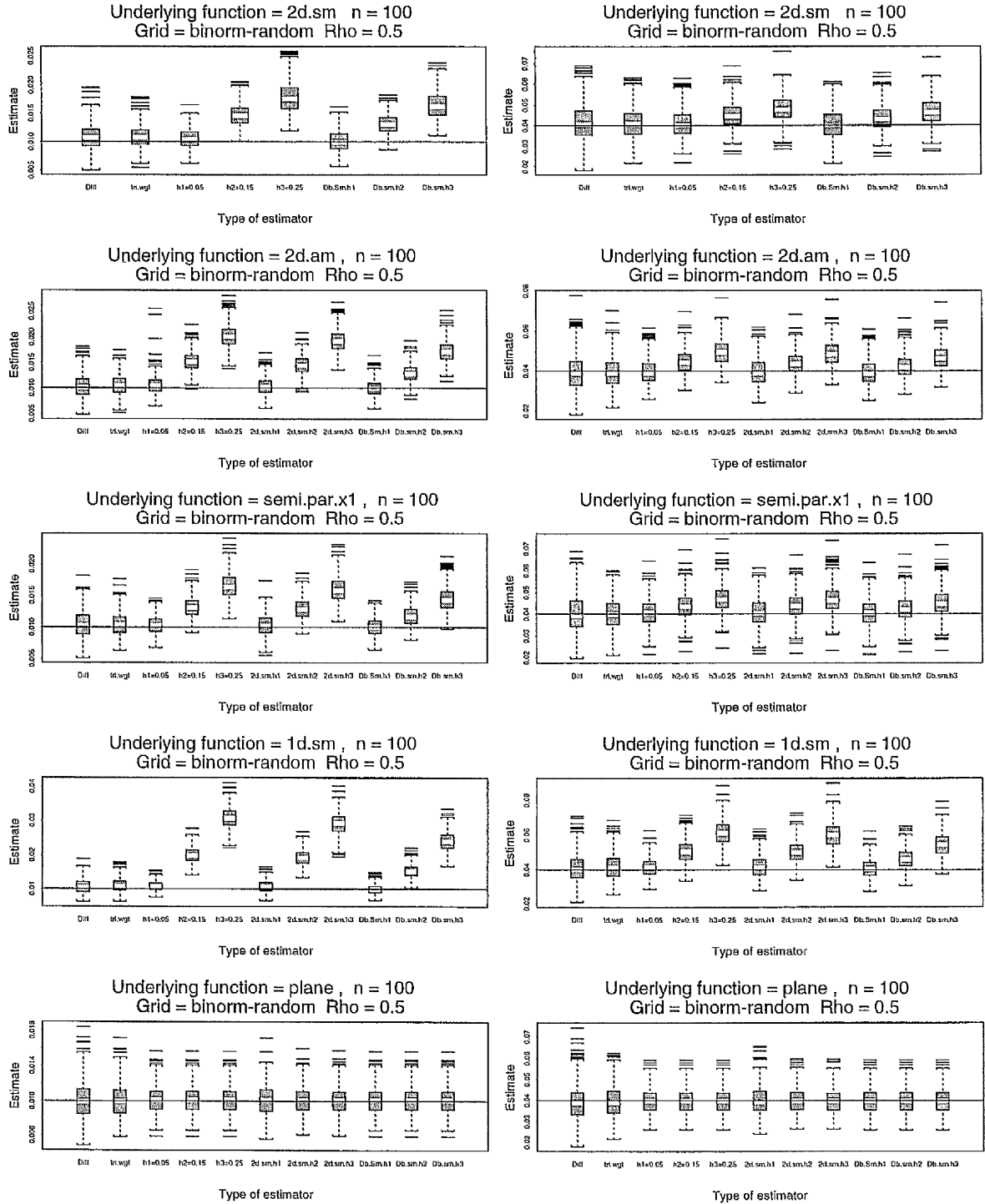
Several modifications of the basic RSS and difference based estimators were also considered. The effect of misspecifying the form of the underlying regression function on the RSS based estimator was investigated by using residuals from



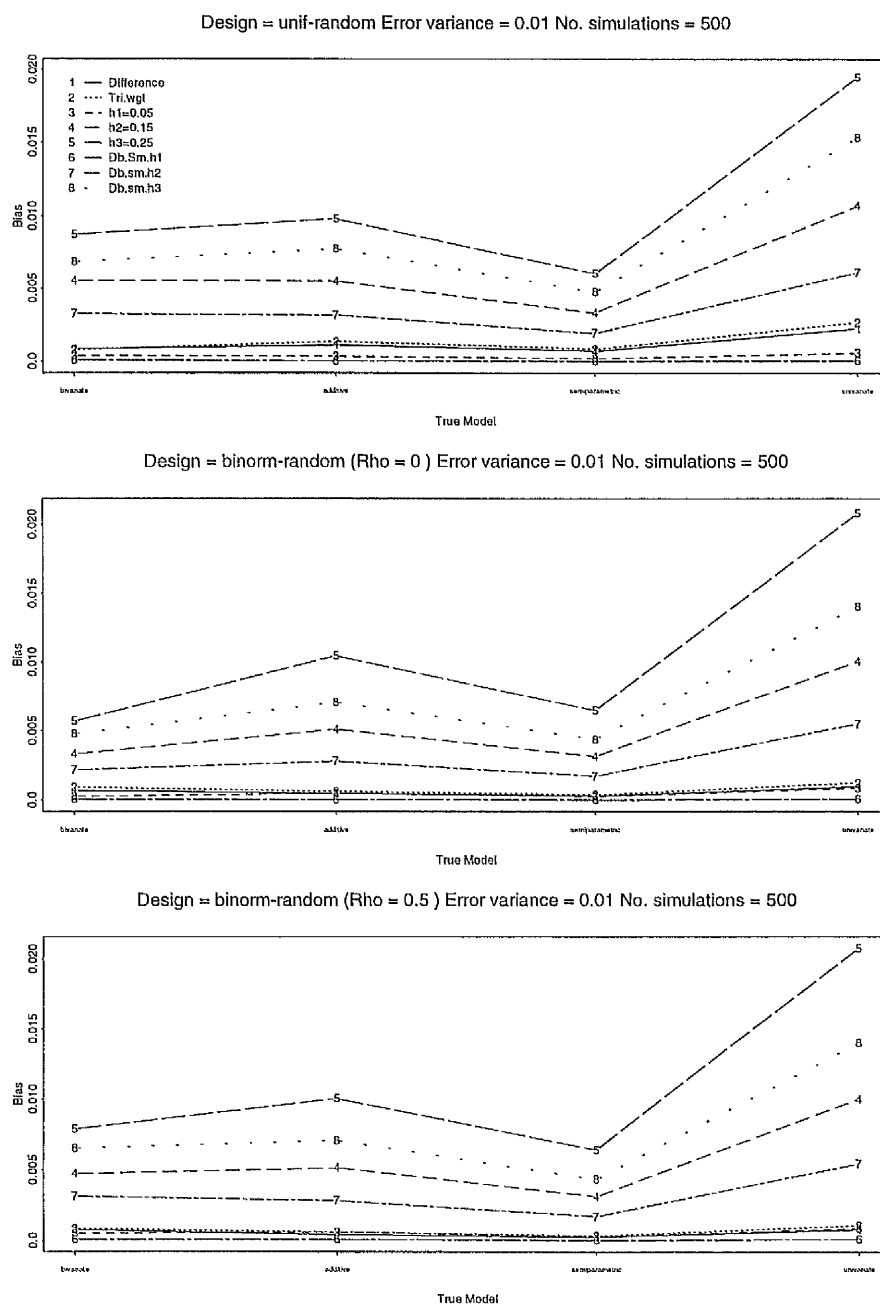
**Figure 2.8.** Boxplots showing the distribution of estimates of  $\sigma^2$  (solid horizontal line) from different estimators. 500 simulations of data generated over uniform-random designs were used.



**Figure 2.9.** Boxplots showing the distribution of estimates of  $\sigma^2$  (solid horizontal line) from different estimators. 500 simulations of data generated over binorm-random ( $\rho = 0$ ) designs were used.



**Figure 2.10.** Boxplots showing the distribution of estimates of  $\sigma^2$  (solid horizontal line) from different estimators. 500 simulations of data generated over binorm-random ( $\rho = 0.5$ ) designs were used.



**Figure 2.11.** Plots showing the finite sample biases of different estimators of  $\sigma^2$  under different regression functions over random designs.



bivariate smooths of the data, regardless of the true form of the regression function. This reflects the practice of using estimates of 'full' model fits to estimate  $\sigma^2$  when nested models are compared (detailed discussion in Chapter 3). The more general model fit yielded results consistent with the true model fit, suggesting that this type of misspecification will not have a major impact on methods of model comparison.

In a random design setting, a modification of the difference based estimator was considered which used weighted triplets of design points to define the pseudoresiduals. This modification did not result in a significant improvement over the bias property of the basic difference based estimator, however. This suggests that the weighting scheme employed in the basic version adapts to different design configurations in an appropriate way.

Having defined and investigated two different approaches to the estimation of  $\sigma^2$ , both of which perform well under certain conditions, it is informative to summarise the ways in which they differ, which in turn suggests alternative estimators. The RSS based and difference based estimators differ with respect to three key concepts: their definitions of *neighbouring points*, their definitions of *residuals* and the *function of the residuals*.

### Defining the neighbours

Perhaps the most striking difference between the RSS approach and the difference based method is the way in which the neighbouring points used to define the 'residual' at each design point are selected.

- ◊ **Delaunay triangulations** were used in the difference based approach as a means of incorporating both the *distance* and *direction* of points relative to one another into the selection of neighbourhoods. Since this approach is a generalisation of univariate difference based techniques, it is natural to carry over the idea of using *surrounding* points to define the pseudoresidual.

This does, however, raise the possibility of incorporating far off points in certain directions. In a univariate context it is easy to imagine the second closest of two points either side of a design point not being the second closest of *all* the points. This is a price of incorporating direction into the definition of neighbours. The weights employed in the calculation of

$\hat{\sigma}^2$  from the pseudoresiduals take this into account, however, and effectively down-weight the influence of far off points. This was confirmed with the use of weighted triplets which again sought to down-weight triplets far from the point of interest. The result that these two methods yielded similar values demonstrates the effectiveness of the simpler weighting scheme.

- ◊ **Bandwidth specification** is at the heart of the RSS based estimator and it is through this and the kernel function used that neighbourhoods are defined. A fixed bandwidth selects neighbours which lie within a fixed region surrounding the design point. As such it does not take into account the direction of neighbours relative to the design points nor does it adapt to changes in the density of the design points.

Alternate approaches, familiar from the nonparametric regression literature, are to use variable bandwidths. For instance a bandwidth at a certain design point is defined to be the distance to the  $k$ th nearest point, or to define *nearest neighbourhoods* to include a fixed *number* of the closest design points. These methods scale the neighbourhoods used to define the residuals, but they still don't explicitly take direction into account.

### Defining the residual

The two approaches differ significantly with respect to the way in which 'residuals' at individual design points are defined.

- ◊ **Interpolation** is used in the difference based method to define pseudoresiduals using the planes through surrounding triplets' responses. Again, this is a generalisation of the familiar univariate technique. However, unlike the univariate case, there is more than one interpolation to consider for each design point and thus the question of how to combine the 'triplet distances' into a single pseudoresidual arises. Both a straight and a weighted average of these distances yielded estimators with similar properties.
- ◊ **Local linear regression** produced the fitted surface, and hence the residuals, used in the RSS based estimator. In this approach there is only one distance to consider since all the information supplied by the neighbouring

points is combined automatically in the locally weighted regression. There is consequently less control over the residuals used and the way they use the local responses compared to the pseudoresidual used in the difference based estimator.

A further contrast with the difference based approach is the rôle of the observation at the design point itself. The RSS approach includes this observation in its estimation of the local surface and thus it appears in both terms whose difference defines the residual at this point. An alternate approach, familiar from the smoothing parameter selection literature, is the use of *cross-validation* which involves leaving the design point out of the estimation of the surface at that point. This is closer to how the difference based estimator uses the response at the design point to calculate the pseudoresidual.

### Form of the estimator

The final difference between the two estimators noted here is the form of the estimator itself, i.e. the way in which the 'residuals' are combined to estimate  $\sigma^2$ .

- ◇ **Individual weights** are applied to each (squared) pseudoresidual in the difference based estimator. It was shown that if the pseudoresiduals had approximately zero means, then the weighting employed made each an unbiased estimate of  $\sigma^2$  and thus their average defines a suitable estimator.
- ◇ **An overall adjustment** applied to the sum of the squared residuals is the approach used by the RSS based estimator. The adjustment comes from the expected value of RSS, and seeks to define an estimator with minimum bias. It therefore assigns an *equal weight* to each residual.

From these comparisons of the two forms of  $\hat{\sigma}^2$ , it is clear that they differ in the way they use individual observations. The difference based estimator takes a very 'micro' approach: defining the neighbouring points in a very local manner (taking into account direction etc.), likewise the pseudoresiduals (interpolating each set of neighbouring triplets individually) and finally weighting individual

pseudoresiduals to define the estimator. In contrast the RSS based approach is a very 'macro' approach: neighbourhoods are fixed by the kernel and applied globally, residuals are defined from an estimate of the global surface (albeit fitted locally) and residuals have equal influence on the final estimate. Combining these 'macro' and 'micro' approaches suggests several alternate approaches to the estimation of  $\sigma^2$ .

The first is to retain the RSS approach but to adjust the estimate of the regression surface used so that it takes into account local features such as the density of design points. One approach is to use a *variable bandwidth* to obtain the estimate of the regression function and hence the residuals. The hope is that the fitted surface, by taking into account local variation in the design points, would reduce the influence of the global bandwidth which was seen to play such a crucial rôle in determining the bias of the RSS based estimator.

However, the issue of bandwidth selection, including the variable versus fixed bandwidth question, is hotly debated in the context of estimating univariate regression functions (see Gasser and Seifert (1994) vs. Cleveland and Loader (1996)). Therefore, to propose an extension to the bivariate setting, with a number of smoothers with different forms of smoothing parameters to choose from, is inviting complexity with little except intuitive justification of an improvement.

For this reason, it seems more profitable to pursue a modification of the difference based approach which will attempt to improve its performance. The idea of incorporating both direction and distance into the definition of a local neighbourhood is an appealing one, especially for the purposes of  $\sigma^2$  estimation, since it mimics most closely the univariate difference based methods. A modification of this method could be to use these 'triangulated' neighbourhoods (i.e. defined by the Delaunay triangulations) but to define a pseudoresidual as the distance between the response and the locally weighted linear fit through the 'triangulated' neighbourhood. This bypasses the question of how best to combine the information in this neighbourhood into a single pseudoresidual since locally weighted regression does this automatically and in a way which transparently down-weights far away points.

Talk of a locally weighted linear fit, however, once again raises the question of what weights and therefore what type of kernel and bandwidth to use. Hence,

we are faced with the same choices and absence of guiding theory as noted above. It requires an extensive and systematic simulation exercise to ascertain which of these local weighting schemes is best suited to error variance estimation. Once the pseudoresiduals are defined, however, estimation of  $\sigma^2$  could proceed as before, i.e. by individually scaling the squared pseudoresiduals and then averaging across all design points.

Although these approaches hold some potential for improving the estimation of  $\sigma^2$  in the bivariate setting, it should be remembered that this task is not a trivial one. Indeed, in the univariate setting there is still considerable uncertainty as to which approach - RSS or difference based - is superior. Indeed, within each of these two established approaches the best methodology is still an open question (see for example Dette *et al.* (1998) for a recent review of some issues).

The complexity of the task must also be weighed against the usefulness of the result. The difference based estimators described here are specific to the bivariate setting and do not generalise easily to higher dimensions. We have restricted attention to bivariate setting since, as the next chapter will show, methods of inference requiring a  $\hat{\sigma}^2$  are suggested by the properties of the model fits. Therefore, we are content to have investigated a single difference based alternative to the RSS estimator and mention the other potential modifications primarily to highlight the issues involved.

The aim of this chapter was not to define an invincible estimator, but rather to raise the neglected issue of the estimation of error variance in the bivariate nonparametric regression setting. The estimation of  $\sigma^2$ , however, is not an end in itself, but rather a means to the 'end' of model inference. The next three chapters will consider the topic of inference via model comparisons. Chapters 3 and 4 will examine the bivariate setting where the material and results of this chapter will be of use. Chapter 5 will extend this to a multi-dimensional setting, with restrictions only on the number of nonparametric components, in which case the only feasible estimator of  $\sigma^2$  is a RSS based one.

## Chapter 3

# Inference Amongst a Class of Bivariate Nonparametric Models: Theoretical Results

### 3.1 Introduction

This chapter develops methods of inference amongst a class of nonparametric regression models. The methods are an extension of tests in the univariate setting, discussed in Section 1.2.3, which compare fits of competing nonparametric models to the same data. They extend these approaches by catering for comparisons amongst a wider class of models, namely a *bivariate class*, which includes model fits such as a bivariate additive fit and a bivariate smooth.

The bivariate class has several features which make it particularly profitable to study. From a methodological viewpoint, explicit definitions of the model fits exist as do theoretical properties such as asymptotic bias. The form of these biases suggest ways in which model comparisons may be made. On a practical level, the span of this class of models is quite broad, capturing a wide variety of forms of the underlying regression function. As such they are quite applicable to real situations where covariate effects are of interest, but little is known about the underlying regression structure. They allow two covariates to appear in various forms ranging from strictly linear to general smooth functions. One area of

applicability is environmental contexts, and other areas where spatial variation is of interest, an example of which is given at the end of Chapter 4. Furthermore these models can be extended through the addition of extra linear terms, as Chapter 5 shows, thus enhancing their potential for modelling real data.

In Section 3.2 the seven models defining this class are described with particular attention given to defining the smoothing matrix which yields the fitted values of each model and an expression for the conditional asymptotic bias of each model. A detailed presentation of the properties of a test statistic based on a comparison of fitted values follows in Section 3.3, including a look at the behaviour of the test statistic under linear models. The distributional properties of the test statistic are described in Section 3.4 which leads naturally to a description of a method of inference for comparing models in Section 3.5.

Chapter 4 summarise the results of an in depth simulation study designed to investigate the properties of the tests developed in this chapter. Two types of design spaces, regular and random, are considered separately and, for each of these, tests are first performed using the known value of the error variance,  $\sigma^2$ , which focuses on the comparisons of model fits only. Secondly, several estimators of  $\sigma^2$ , described in Chapter 2, are employed to define and assess model comparison procedures useful in practice.

## 3.2 Models Under Consideration

### 3.2.1 Defining the fitted values and bias of each of the models

Consider samples of  $n$  triplets of variables  $(X, Z, Y)$ ,  $\mathbf{x} = (X_1, \dots, X_n)^T$ ,  $\mathbf{z} = (Z_1, \dots, Z_n)^T$  and  $\mathbf{y} = (Y_1, \dots, Y_n)^T$ , where  $\mathbf{x}$ ,  $\mathbf{z}$  and  $\mathbf{y}$  are related by

$$\mathbf{y} = m(\mathbf{x}, \mathbf{z}) + \boldsymbol{\varepsilon}$$

where  $m(\mathbf{x}, \mathbf{z}) = \{m(X_1, Z_1), \dots, m(X_n, Z_n)\}$  for a smooth bivariate regression function  $m(\cdot, \cdot)$  and  $\boldsymbol{\varepsilon}$  an  $n$ -vector of independent zero mean normal errors  $\varepsilon_i$  with  $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$ , where  $\mathbf{I}$  is the  $n \times n$  identity matrix and  $\sigma^2$  is the unknown

**Table 3.1.** Seven models defining the bivariate class

Model	Abbreviation	Regression function
no-effect (constant)	const.	$\mu$
simple linear	st.line	$\mu + \beta X$
bivariate linear	plane	$\mu + \beta_1 X + \beta_2 Z$
univariate smooth	1d.sm	$\mu + m_1(X)$
semiparametric	semi.par	$\mu + \beta X + m_1(Z)$
bivariate additive	2d.am	$\mu + m_1(X) + m_2(Z)$
bivariate smooth	2d.sm	$\mu + m^*(X, Z)$

(constant) error variance.

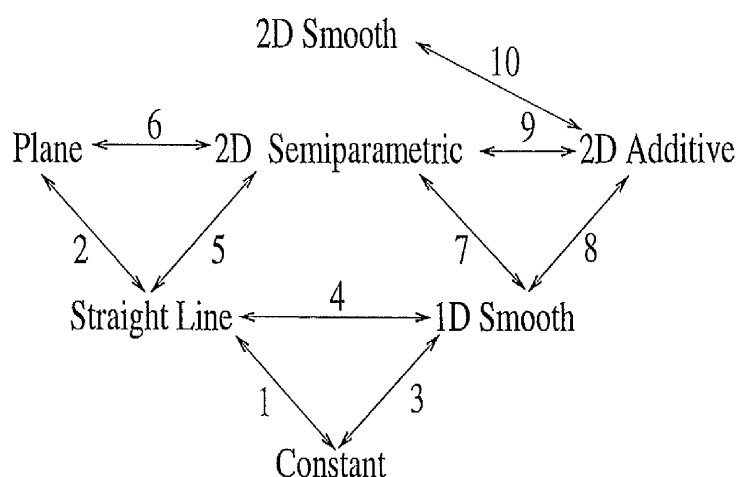
Table 3.1 lists seven forms of  $m(X, Z)$  which define the bivariate class of models considered in this chapter. The table also introduces the abbreviations and notation<sup>1</sup> to be used henceforth. Note the defining qualities of models in this class:

- ◊ at most, they involve two covariates;
- ◊ when two terms appear in the model they combine in an additive fashion;
- ◊ covariates can appear as either a linear term (with an unknown parameter) or as the argument(s) of an (unknown) smooth function.

Figure 3.1 also lists the seven models, this time reflecting the increasing complexity and nested nature of the models. At the top level is the bivariate smooth model which is the most general way two covariates can define the underlying regression surface. The second row lists the three models which also are in terms of two covariates but these appear as separate terms, either both linear (plane) or both as arguments to smooth functions (additive) or one of each (semiparametric). The third row lists the two possible univariate models which assume that there is no effect in the direction of the second covariate. Finally the bottom row consists of the ‘constant’ model which allows for the possibility of no effect in either direction.

<sup>1</sup>Although the regression function  $m(\cdot)$  is defined generally to have two arguments, we will subsequently drop redundant variables.





**Figure 3.1.** Diagram showing the hierarchy of models and natural model comparisons amongst the bivariate class of models.

The lines in Figure 3.1 link models which differ by either one or two degrees of complexity. As such they suggest ways in which inference may proceed by comparing model fits of connected types to assess the effect of covariates. Comparisons 1 and 2 proceed using standard linear model inference. Comparisons 3 and 4 can be made using univariate nonparametric model inference techniques described in Section 1.2.3. The remaining comparisons have not received specific attention in the research literature. The aim of this chapter is to derive methods to perform these comparisons. Table 1.7 lists these comparisons<sup>2</sup>, introducing the terminology ‘reduced’ and ‘full’ models used throughout the chapter.

The method of estimation used to obtain the nonparametric fit is local linear regression (see Fan & Gijbels (1996)). The local linear method has the benefit of several excellent theoretical properties (described in Chapter 1) as well as having the intuitive advantage of being a relaxation of the usual linear regression model. It is also a linear smoother (as are splines and kernel smoothers), i.e. a vector of fitted values can be defined as  $\hat{m}(\mathbf{x}, \mathbf{z}) \equiv \hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$  for a suitable  $n \times n$  smoothing matrix  $\mathbf{S}$ . For each of the models listed in Table 3.1 the fitted values are defined in the following sections by specifying the smoothing matrix  $\mathbf{S}$  which yields the

<sup>2</sup>Comparisons 5 and 6 are similar in that they both compare a linear model fit with a nonparametric model and therefore only one (comparison 5) is included here.

**Table 3.2.** Comparisons amongst the bivariate class of models considered in the simulation studies of Chapter 4

Comparison	Reduced Model	Full Model
2d.am vs. 2d.sm	$\mu + m_1(X) + m_2(Z)$	$\mu + m(X, Z)$
semi.par.x1 vs. 2d.am	$\mu + X\beta + m(Z)$	$\mu + m_1(X) + m_2(Z)$
1d.sm vs. 2d.am	$\mu + m(X)$	$\mu + m_1(X) + m_2(Z)$
1d.sm vs. semi.par.x1	$\mu + m(X)$	$\mu + m(X) + \beta Z$
st.line vs. semi.par.x2	$\mu + X\beta$	$\mu + \beta X + m(Z)$

fitted values. Expressions for the asymptotic bias of each model fit are also listed, the common form of which suggests how the model fits could best be compared.

### 3.2.2 Constant (no-effect) model

For the trivial case,

$$E(\mathbf{y}|X = \mathbf{x}, Z = \mathbf{z}) \equiv m(\mathbf{x}) = \boldsymbol{\mu} ,$$

$\boldsymbol{\mu}$  a constant  $n$ -vector, the smoothing matrix is defined as

$$\mathbf{S} = \frac{1}{n} \mathbf{1} \mathbf{1}^T$$

where  $\mathbf{1}$  is an  $n$ -vector of 1's.

Clearly each element of  $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$  is an unbiased estimate of the mean  $\mu$ .

### 3.2.3 Simple and bivariate linear regression model

For the familiar linear parametric models,

$$E(\mathbf{y}|X = \mathbf{x}, Z = \mathbf{z}) \equiv m(\mathbf{x}) = \boldsymbol{\mu} + \mathbf{x}\beta_1$$

$$E(\mathbf{y}|X = \mathbf{x}, Z = \mathbf{z}) \equiv m(\mathbf{x}, \mathbf{z}) = \boldsymbol{\mu} + \mathbf{x}\beta_1 + \mathbf{z}\beta_2,$$

linear regression results define the smoothing (hat) matrix as,

$$\mathbf{S} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T.$$

In the case of simple linear regression  $\mathbf{X} = (1, \mathbf{x})$  and for bivariate regression  $\mathbf{X} = (1, \mathbf{x}, \mathbf{z})$ . Standard results show that both of these model fits are unbiased estimates of the true (unknown) mean responses.

### 3.2.4 Univariate smooth model

The univariate nonparametric model has the form,

$$E(\mathbf{y}|X = \mathbf{x}, Z = \mathbf{z}) \equiv m(\mathbf{x}) = \boldsymbol{\mu} + m_1(\mathbf{x}),$$

for some univariate smooth function  $m_1(\cdot)$ . The local linear regression smoother was introduced briefly in Section 1.2.1. Here we detail the method using the matrix notation introduced above.

To calculate the local linear estimate of the regression function at an arbitrary value of  $X$  in the domain of  $m(\cdot)$ ,  $x_0$  say, first define a *design matrix*:

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - x_0) \\ 1 & (X_2 - x_0) \\ \vdots & \vdots \\ 1 & (X_n - x_0) \end{pmatrix}.$$

Also define  $\mathbf{W}$  to be the  $n \times n$  diagonal matrix of weights:

$$\mathbf{W} = \text{diag}\{K_h(X_i - x_0)\},$$

where  $K_h(z) = K(z/h)/h$  is the *kernel function*. Throughout this and subsequent chapters we will use the standard normal density function for  $K(\cdot)$ .

The *local linear regression* estimate of  $m(x_0)$  is given by the component  $\hat{\beta}_0$  of

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}. \quad (3.1)$$

To construct an estimate of  $m(\cdot)$  over the design space of  $X$  the above procedure would be repeated for different values of  $x_0$ . Let  $\mathbf{x}_{\text{eval}} = (x_{01}, x_{02}, \dots, x_{0n'})$  be  $n'$  such *evaluation points*. Unless otherwise stated, we will use  $\mathbf{x}_{\text{eval}} = \mathbf{x}$ , the vector of observed design points, but the notation emphasises that the estimate of  $m(\cdot)$  can be calculated at any set of points in the domain.

A consequence of the *linear* nature of local polynomial regression smoothers is that fitted values can always be defined as a linear combination of the observed responses. The local linear fit described above can be defined explicitly as  $\hat{m}(x_0) = \sum_{j=1}^n w_j(x_0)Y_j$  using weights

$$w_j(x_0) = \frac{(s_2(x_0; h) - s_1(x_0; h)(X_j - x_0))K((X_j - x_0)/h)}{n(s_2(x_0; h)s_0(x_0; h) - s_1(x_0; h)^2)}$$

where  $s_r(x_0; h) = \sum_{j=1}^n (X_j - x_0)^r K((X_j - x_0)/h)/n$  (see Bowman and Azzalini, 1997, pp. 50).

It follows that the local linear *smoothing matrix*  $\mathbf{S}$  which will yield estimates of  $m(\cdot)$  at the *observed*  $X$  values, has an  $(i, j)$ th element given by  $s_{ij} = w_j(X_i)$ . When applied to the vector of observed responses, this smoothing matrix yields a  $n$ -vector of the fitted values at the design points, i.e.

$$\hat{\mathbf{y}} \equiv \hat{\mathbf{m}}(\mathbf{x}) = \mathbf{S}\mathbf{y}.$$

We shall see in later sections that it is sometimes necessary to add constraints to the smoothing technique used. For instance, to ensure the identifiability of higher dimensional models the univariate smoother must return fitted values having mean zero, i.e.  $\sum_i \hat{m}(X_i) = 0$ . This is achieved via a *centred* version of  $\mathbf{S}$ ,

$$\mathbf{S}_c = (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{S}.$$

This yields an estimator of the nonparametric component  $m_1(\cdot)$ , i.e.  $\hat{m}_1(\mathbf{x}) = \mathbf{S}_c\mathbf{y}$ . If we also define the  $n$ -vector  $\bar{\mathbf{y}} = \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{y}$ , the *centred* fitted values become,

$$\hat{m}_c(\mathbf{x}) \equiv \hat{\mathbf{y}}_c = \bar{\mathbf{y}} + \mathbf{S}_c\mathbf{y} = (\frac{1}{n}\mathbf{1}\mathbf{1}^T + \mathbf{S}_c)\mathbf{y} = (\frac{1}{n}\mathbf{1}\mathbf{1}^T(\mathbf{I} - \mathbf{S}) + \mathbf{S})\mathbf{y}.$$

This notation shows that using a centred smoothing matrix ensures that the residuals from the fit have zero mean.

It has been shown (e.g. Wand and Jones, 1995) that using the (uncentred) local linear smooth, i.e.  $\hat{m}(\mathbf{y}) = \mathbf{S}\mathbf{y}$ , produces an estimate with asymptotic conditional bias at a particular point  $x_0$  given by:

$$E(\hat{m}(x_0) - m(x_0)|\mathbf{x}) = \frac{h^2}{2}\mu_2(K)m''(x_0) + o_P(h^2)$$

where  $\mu_2(K) = \int u^2 K(u)du$ , i.e. the variance of the kernel function.

Similarly for the *centred* fit,  $\hat{m}_c(\mathbf{y}) = (\frac{1}{n}\mathbf{1}\mathbf{1}^T - \mathbf{S}_c)\mathbf{y}$ , the asymptotic conditional bias can be approximated as follows,

$$\begin{aligned} E(\hat{m}_c(x_0) - m(x_0)|\mathbf{x}) &= E\left(\frac{1}{n}\sum_{i=1}^n Y_i + \hat{m}(x_0) - \frac{1}{n}\sum_{i=1}^n \hat{m}(X_i) - m(x_0)|\mathbf{x}\right) \\ &= \underbrace{E(\hat{m}(x_0)|\mathbf{x}) - m(x_0)}_{\text{uncentred bias}} - \frac{1}{n}\sum_{i=1}^n \underbrace{(E(\hat{m}(X_i)|\mathbf{x}) - m(X_i))}_{\text{uncentred bias}} \\ &\approx \frac{h^2}{2}\mu_2(K)(m''(x) - E(m''(\cdot))) \end{aligned}$$

### 3.2.5 Semiparametric model

The semiparametric model with a single linear covariate  $X$  and a single nonparametric component in  $Z$  has the form:

$$E(\mathbf{y}|X = \mathbf{x}, Z = \mathbf{z}) \equiv m(\mathbf{x}, \mathbf{z}) = \mathbf{X}\boldsymbol{\beta} + m_1(\mathbf{z}),$$

where the design matrix and the parameter vector are respectively:

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 - \bar{X} \\ 1 & X_2 - \bar{X} \\ \vdots & \vdots \\ 1 & X_n - \bar{X} \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

Although we are focusing on the bivariate class of models, extensions to more than one covariate appearing linearly are possible. We shall consider this extension in Chapter 5. Also, the inclusion of an intercept term in the parametric component is intended to aid comparisons with other models which require that the overall mean is estimated separately, i.e. not included in the nonparametric component.

Within the semiparametric model there are two components, one linear and one nonparametric, which need to be estimated. One set of solutions, derived by analogy with the backfitting algorithm (see Hastie and Tibshirani (1990) pp. 118) uses least squares to estimate the parameters and a smoother to estimate the nonparametric component. Let  $\mathbf{S}_z$  be the *centred smoothing matrix*<sup>3</sup> derived using local linear regression based on the observed values  $\mathbf{z}$ , then the backfitting algorithm has explicit solutions:

$$\hat{\beta} = (\mathbf{X}^T(\mathbf{I} - \mathbf{S}_z)\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{S}_z)\mathbf{y} \quad (3.2)$$

$$\hat{m}_1(\mathbf{z}) = \mathbf{S}_z(\mathbf{y} - \mathbf{X}\hat{\beta}). \quad (3.3)$$

Substituting these definitions into the model gives a smoothing matrix which yields the fitted values  $\hat{m}(\mathbf{x}, \mathbf{z}) = \mathbf{S}\mathbf{y}$  where

$$\mathbf{S} = \mathbf{X}(\mathbf{X}^T(\mathbf{I} - \mathbf{S}_z)\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{S}_z) + \mathbf{S}_z(\mathbf{I} - \mathbf{X}(\mathbf{X}^T(\mathbf{I} - \mathbf{S}_z)\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{S}_z))$$

The conditional asymptotic bias of the semiparametric model fit (when the true regression function has the semiparametric form) is given by:

$$\begin{aligned} E(\hat{m}(\mathbf{x}, \mathbf{z}) - m(\mathbf{x}, \mathbf{z})|\mathbf{x}) &= E(\mathbf{X}\hat{\beta} + \hat{m}_1(\mathbf{z})) - \mathbf{X}\beta - m(\mathbf{z}) \\ &= E(\mathbf{X}\hat{\beta} + \mathbf{S}_z(\mathbf{y} - \mathbf{X}\hat{\beta})) - \mathbf{X}\beta - m(\mathbf{z}) \\ &= \mathbf{X}E(\hat{\beta}) + \mathbf{S}_z(\mathbf{X}\beta + m(\mathbf{z}) - \mathbf{X}E(\hat{\beta})) - \mathbf{X}\beta - m(\mathbf{z}) \end{aligned}$$

---

<sup>3</sup>It has been noted (Opsomer and Ruppert, 1999 p4) that the use of centred smoothers is one solution to the often overlooked fact that the estimators are *not* well-defined when one of the parametric terms is taken to be an estimate of the overall mean.

$$\begin{aligned}
&= (\mathbf{I} - \mathbf{S}_z)\mathbf{X}E(\hat{\beta}) - (\mathbf{I} - \mathbf{S}_z)\mathbf{X}\beta + \mathbf{S}_zm(\mathbf{z}) - m(\mathbf{z}) \\
&= (\mathbf{I} - \mathbf{S}_z)\mathbf{X} \underbrace{(E(\hat{\beta}) - \beta)}_1 + \underbrace{\mathbf{S}_zm(\mathbf{z}) - m(\mathbf{z})}_2.
\end{aligned}$$

The two components marked 1 and 2 are the biases of the parameter estimates and the nonparametric component estimate respectively.

The parametric bias component (1) has been the focus of considerable attention over the years. Rice (1986) showed that the bias of the estimator  $\hat{\beta}$  can asymptotically dominate the variance when the covariates  $X$  and  $Z$  are correlated. This prompted Speckman (1988) to suggest replacing  $(\mathbf{I} - \mathbf{S}_z)$  in the definition of  $\hat{\beta}$  with  $(\mathbf{I} - \mathbf{S}_z)^T(\mathbf{I} - \mathbf{S}_z)$  although this treatment was in the context of kernel (local constant) regression. We shall consider this estimator further in Section 5.2.1. Speckman's estimates are not solutions to the backfitting algorithm, however, nor do they generalise easily to models with more than one nonparametric component (subject of Chapter 5). Therefore we will restrict attention here to the *backfitting estimates* defined above.

Opsomer and Ruppert (1999) recently considered this model and these backfitting solutions in the context of local linear regression. They show that under suitable assumptions, the asymptotic conditional bias<sup>4</sup> of the parameter estimates is given by:

$$E(\hat{\beta} - \beta | \mathbf{x}, \mathbf{z}) = -\frac{\mu_2(K)}{2}h^2 E(\text{var}(X_i|Z_i))^{-1} \text{cov}(X_i, m''(Z_i)) + o_p(h^2) \quad (3.4)$$

Note that the bias of the parameter estimate is a function of the joint distribution of  $X_i$  and  $Z_i$  through the terms  $\text{cov}(X_i, m''(Z_i))$  and  $E(\text{var}(X_i|Z_i))$ . When the covariates are independent these terms are zero and the estimate of the parametric component is unbiased. One immediate implication of this is that the estimate of the intercept (overall mean) is unbiased since the first column of  $\mathbf{X}$  consists of 1's and thus has no association with the  $Z_i$ 's.

---

<sup>4</sup>They also use these results to define a smoothing parameter selection method which ensures that the backfitting estimator of  $\beta$  has convergence properties equivalent to Speckman's estimator based on cross validation.

The second component is the familiar bias of a *centred* univariate smooth, defined in Section 3.2.4. Since there is nothing in this second component involving  $\hat{\beta}$  or  $\mathbf{X}$ , its behaviour (including bias) is identical to the univariate case.

Therefore, the conditional asymptotic bias of the semiparametric model fit defined by Equations 3.2 is

$$E(\hat{m}(x_0, z_0) - m(x_0, z_0) | \mathbf{x}, \mathbf{z}) \approx \frac{h^2}{2} \mu_2(K) m_1''(z_0) - \frac{h^2}{2} \mu_2(K) E(m_1''(\cdot)) - \frac{\mu_2(K)}{2} h^2 E(\text{var}(X_i | Z_i))^{-1} \text{cov}(X_i, m_1''(Z_i))$$

and when  $X$  and  $Z$  are independent this reduces to:

$$E(\hat{m}(x_0, z_0) - m(x_0, z_0) | \mathbf{x}, \mathbf{z}) \approx \frac{h^2}{2} \mu_2(K) (m_1''(z_0) - E(m_1''(\cdot)))$$

The use of a centred smoothing matrix in the definition of the fitted values is evident by the presence of the expectation of the second derivative over the design space.

### 3.2.6 Additive model

The additive model with two covariates  $X$  and  $Z$  has the form:

$$E(\mathbf{y} | X = \mathbf{x}, Z = \mathbf{z}) = m(\mathbf{x}, \mathbf{z}) = \mu + m_1(\mathbf{x}) + m_2(\mathbf{z}).$$

Opsomer and Ruppert (1997) study the estimators of  $m_1(\cdot)$  and  $m_2(\cdot)$  when the bivariate additive model is fit using local polynomial (and thus linear) regression.

As noted earlier (Section 1.3.2) the iterative backfitting algorithm, which is the usual method of fitting additive models of any dimension, yields explicit solutions in the bivariate case. The smoothing matrix which yields the fitted bivariate additive model can be defined as

$$\mathbf{S} = \frac{1}{n} \mathbf{1}\mathbf{1}^T + 2\mathbf{I} - (\mathbf{I} - \mathbf{S}_x \mathbf{S}_z)^{-1} (\mathbf{I} - \mathbf{S}_x) - (\mathbf{I} - \mathbf{S}_x \mathbf{S}_z)^{-1} (\mathbf{I} - \mathbf{S}_z)$$

where  $\mathbf{S}_x$  and  $\mathbf{S}_z$  are the *centred* univariate smoothing matrices associated with



$\mathbf{x}$  and  $\mathbf{z}$  respectively, based on kernels  $K_1$  and  $K_2$  and smoothing parameters  $h_1$  and  $h_2$  respectively. In this setting, as in the semiparametric model, it is *necessary* to employ centred smoothers to ensure the uniqueness of the solutions. In other words, employing centred smoothers ensures the existence of the matrix  $(\mathbf{I} - \mathbf{S}_x \mathbf{S}_z)^{-1}$  in the definition of  $\mathbf{S}$ .

Often it is of interest to estimate the two components of the additive model,  $m_1(\cdot)$  and  $m_2(\cdot)$ , separately. These fits are given by

$$\hat{m}_1(\mathbf{x}) = \{\mathbf{I} - (\mathbf{I} - \mathbf{S}_x \mathbf{S}_z)^{-1}(\mathbf{I} - \mathbf{S}_x)\}\mathbf{y} \quad \text{and}$$

$$\hat{m}_2(\mathbf{z}) = \{\mathbf{I} - (\mathbf{I} - \mathbf{S}_z \mathbf{S}_x)^{-1}(\mathbf{I} - \mathbf{S}_z)\}\mathbf{y}.$$

The expression for the conditional asymptotic bias of a fitted additive model is also derived and discussed by Opsomer and Ruppert (1997). Different expressions apply depending on the location in the design space, namely for interior and boundary points. The joint and marginal distributions of the covariates,  $f(X, Z)$ ,  $f_X(X)$  and  $f_Z(Z)$ , also enter the expression in several places. Information can be summarised for the observed design points in the form of the  $n \times n$  matrix  $\mathbf{T}$  whose  $(i, j)$ th element is

$$T_{ij} = \frac{1}{n} \frac{f(X_i, Z_j)}{f_X(X_i)f_Z(Z_j)} - \frac{1}{n}.$$

Let the  $i$ th row and the  $j$ th column of  $(\mathbf{I} - \mathbf{T})^{-1}$  be denoted by  $\mathbf{t}_i^T$  and  $\mathbf{v}_j$  respectively. If the observed point  $(X_i, Z_i)$  lies in the interior of  $\text{supp}(f)$ , the bias of  $\hat{m}(X_i, Z_i)$  is given by

$$\begin{aligned} E(\hat{m}(X_i, Z_i) - m(X_i, Z_i) | \mathbf{x}, \mathbf{z}) &= \frac{1}{2} h_1^2 \mu_2(K_1) (\mathbf{t}_i^T \mathbf{D}^2 \mathbf{m}_1 - \mathbf{v}_i^T E(m_1''(X_i) | \mathbf{z})) + \\ &\quad \frac{1}{2} h_2^2 \mu_2(K_2) (\mathbf{v}_i^T \mathbf{D}^2 \mathbf{m}_2 - \mathbf{t}_i^T E(m_2''(Z_i) | \mathbf{z})) \\ &\quad + o_p(h_1^2 + h_2^2), \end{aligned}$$

where

$$\mathbf{D}^2 \mathbf{m}_1 = \begin{bmatrix} \frac{d^2 m_1(X_1)}{dx^2} \\ \vdots \\ \frac{d^2 m_1(X_n)}{dx^2} \end{bmatrix}$$

and

$$E(m_1''(X_i|\mathbf{z})) = \begin{bmatrix} E(m_1''(X_i|Z_1)) \\ \vdots \\ E(m_1''(X_i|Z_n)) \end{bmatrix}$$

and analogously for  $\mathbf{D}^2 \mathbf{m}_2$  and  $E(m_2''(Z_i|\mathbf{x}))$ .

If however we assume that the covariates are statistically independent (i.e. that their joint distribution is the product of the marginal distributions), then the expression simplifies considerably to:

$$\begin{aligned} E(\hat{m}(X_i, Z_i) - m(X_i, Z_i)|\mathbf{x}, \mathbf{z}) &= \frac{h_1^2}{2} \mu_2(K_1)(m_1''(X_i) - E(m_1''(\cdot))) + \\ &\quad \frac{h_2^2}{2} \mu_2(K_2)(m_2''(Z_i) - E(m_2''(\cdot))) \\ &\quad + o_p(h_1^2 + h_2^2) \end{aligned}$$

Once again, the presence of centred smoothing matrices in the definition yield the expectation terms in the expression for bias. Note also the allowance for different kernel and smoothing parameters in the two covariates.

### 3.2.7 Bivariate smooth model

The bivariate smooth model has the form:

$$E(\mathbf{y}|X = \mathbf{x}, Z = \mathbf{z}) \equiv m(\mathbf{x}, \mathbf{z}) = \boldsymbol{\mu} + m^*(\mathbf{x}, \mathbf{z}).$$

The fitted values of a local linear bivariate smooth are obtained by a simple extension of the univariate local linear smooth. In this setting a local *plane* is estimated at each evaluation point, and its fit at that point is used as the bivariate smooth estimate of the regression function. At a point in the design

space,  $(x_0, z_0)$  say, this fit is given by

$$\hat{m}(x_0, z_0) = [1, 0, 0](\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}$$

where  $\mathbf{X}$  is an  $n \times 3$  design matrix whose  $i$ th row is  $(1, X_i - x_0, Z_i - z_0)$  and  $\mathbf{W}$  contains the normal product kernels, i.e. an  $n \times n$  matrix with  $K_2(x_0, z_0) = K((X_i - x_0)/h_1)K((Z_i - z_0)/h_2)$  on the diagonal and zeroes elsewhere<sup>5</sup>. This indicates how the complete smoothing matrix  $\mathbf{S}$  would be constructed to yield  $\hat{m}(\mathbf{x}, \mathbf{z}) = \mathbf{S} \mathbf{y}$ .

The conditional asymptotic bias of this fit was shown by Ruppert and Wand (1994) to be:

$$E(\hat{m}(X_i, Z_i) - m(X_i, Z_i) | \mathbf{x}, \mathbf{z}) = \frac{1}{2} \mu_2(K) \left( h_1^2 \frac{\partial^2 m}{\partial X^2} \Big|_{X_i, Z_i} + h_2^2 \frac{\partial^2 m}{\partial Z^2} \Big|_{X_i, Z_i} \right).$$

Because the kernel function  $K_2(\cdot, \cdot)$  is a bivariate function, a slightly different definition of  $\mu_2(K)$  to that given in Section 3.2.4 is needed. An assumption underlying this asymptotic result is that there exists a scalar,  $\mu_2(K)$  which satisfies:

$$\int \int \begin{bmatrix} x^2 & xz \\ xz & z^2 \end{bmatrix} K_2(x, z) dx dz = \begin{bmatrix} \mu_2(K) & 0 \\ 0 & \mu_2(K) \end{bmatrix}.$$

The normal product kernel satisfies this condition in a straight forward manner.

As in the univariate case, a centred version of the bivariate smooth model can easily be constructed as

$$\mathbf{S}^* = (\mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^T) \mathbf{S}$$

which would return an estimate  $\hat{m}^*(\mathbf{x}, \mathbf{z}) = \mathbf{S}^* \mathbf{y}$  of  $m^*(\mathbf{x}, \mathbf{z})$ . Combining this with the estimate of the overall mean,  $\bar{\mathbf{y}}$ , yields the centred fit

$$\hat{m}_c(\mathbf{x}, \mathbf{z}) = (\frac{1}{n} \mathbf{1} \mathbf{1}^T + \mathbf{S}^*) \mathbf{y} = (\frac{1}{n} \mathbf{1} \mathbf{1}^T (\mathbf{I} - \mathbf{S}) + \mathbf{S}) \mathbf{y}$$

---

<sup>5</sup>This 'product kernel' is a natural choice for a bivariate kernel function.

Using this estimator an expression for the bias can be derived as follows,

$$\begin{aligned}
 E(\hat{m}_c(X_i, Z_i) - m(X_i, Z_i) | \mathbf{x}, \mathbf{z}) &= E(\bar{Y} + \hat{m}(X_i, Z_i) - \frac{1}{n} \sum_{i=1}^n \hat{m}(X_i, Z_i) - \\
 &\quad m(X_i, Z_i) | \mathbf{x}, \mathbf{z}) \\
 &= \underbrace{E(\hat{m}(X_i, Z_i)) - m(X_i, Z_i)}_{\text{uncentred bias}} - \\
 &\quad \frac{1}{n} \sum_{i=1}^n \underbrace{(E(\hat{m}(X_i, Z_i)) - m(X_i, Z_i))}_{\text{uncentred bias}} \\
 &\approx \frac{1}{2} \mu_2(K) \left( h_1^2 \left\{ \frac{\partial^2 m}{\partial X^2} \Big|_{X_i, Z_i} - E \left( \frac{\partial^2 m}{\partial X^2} \right) \right\} + \right. \\
 &\quad \left. h_2^2 \left\{ \frac{\partial^2 m}{\partial Z^2} \Big|_{X_i, Z_i} - E \left( \frac{\partial^2 m}{\partial Z^2} \right) \right\} \right)
 \end{aligned}$$

### 3.2.8 Summary

The asymptotic bias results for the centred model fits discussed in this section are summarised here. Note the similar form of these expressions.

- |                              |          |
|------------------------------|----------|
| 1: Constant:                 | unbiased |
| 2: Simple linear:            | unbiased |
| 3: Planar:                   | unbiased |
| 4: Univariate nonparametric: |          |

$$\frac{h_1^2}{2} \mu_2(K) \{m_1''(x_0) - E(m_1''(\cdot))\}$$

- 5: Semiparametric (independent covariates):

$$\frac{h_1^2}{2} \mu_2(K) \{m_1''(x_0) - E(m_1''(\cdot))\}$$

- 6: Additive model (independent covariates):

$$\frac{h_1^2}{2} \mu_2(K_1) \{m_1''(x_0) - E(m_1''(\cdot))\} + \frac{h_2^2}{2} \mu_2(K_2) \{m_2''(z_0) - E(m_2''(\cdot))\}$$

**Table 3.3.** Asymptotic bias expressions of 'full' model fits when the regression function is of a 'reduced' form

Regression Function	Model Fit	Bias of Model Fit
st.line	semi.par	unbiased
1d.sm: $\mu + m_1(X)$	semi.par	$\frac{h_1^2}{2}\mu_2(K)\{m_1''(x_0) - E(m_1''(\cdot))\}$
1d.sm: $\mu + m_1(X)$	2d.am	$\frac{h_1^2}{2}\mu_2(K_1)\{m_1''(x_0) - E(m_1''(\cdot))\}$
semi.par $\mu + X\beta + m_1(Z)$	2d.am	$\frac{h_1^2}{2}\mu_2(K_1)\{m_1''(z_0) - E(m_1''(\cdot))\}$
2d.am: $\mu + m_1(X) + m_2(Z)$	2d.sm	$\frac{h_1^2}{2}\mu_2(K_1)\{m_1''(x_0) - E(m_1''(\cdot))\} +$ $\frac{h_2^2}{2}\mu_2(K_2)\{m_2''(z_0) - E(m_2''(\cdot))\}$

7: Bivariate nonparametric:

$$\frac{h_1^2}{2}\mu_2(K)\left\{\left.\frac{\partial^2 m}{\partial X^2}\right|_{x_0, z_0} - E\left(\frac{\partial^2 m}{\partial X^2}\right)\right\} + \frac{h_2^2}{2}\mu_2(K)\left\{\left.\frac{\partial^2 m}{\partial Z^2}\right|_{x_0, z_0} - E\left(\frac{\partial^2 m}{\partial Z^2}\right)\right\}$$

The rest of this chapter develops tests based on the comparisons of model fits to perform model inference. Since bias plays a dominant role in these methods, it is instructive to outline here the bias expressions listed above in this context.

Table 3.2 listed the comparisons of model fits which define a systematic approach to assessing covariate effects. Of particular interest for each comparison is the form of the bias expression of the more general (*full*) model when the regression function has the simpler (*reduced*) form. These asymptotic biases are listed in Table 3.3 and show that in each case, when the covariates are independently distributed, the asymptotic bias of the full and reduced (true) model fits are identical provided that equivalent *smoothing parameters* and *kernel functions* are employed in the two fits.

If the covariates are *not* independent then extra terms appear in the expressions for the bias of the semiparametric and additive fits. If the true regression function has the form  $\mu + m(X)$  (univariate smooth) then semiparametric and additive fits have approximate conditional asymptotic bias shown in Table 3.4. This same expression for the bias of the additive model also applies when the regression function has a semiparametric form, since the second derivative of the

**Table 3.4.** Asymptotic bias expressions of semiparametric and additive model fits when the regression function is a univariate function over dependent covariates.

Model Fit	Bias of Fit for Univariate Reg. Fn.
semi.par	$\frac{h^2}{2}\mu_2(K)\{m''(X_i) - E(m''(\cdot)) - \frac{\text{COV}(X_i, m''(Z_i))}{E(\text{Var}(X_i Z_i))}\}$
2d.am	$\frac{1}{2}h_1^2\mu_2(K_1)(\mathbf{t}_i^T \mathbf{D}^2 \mathbf{m}_1 - \mathbf{v}_i^T E(m_1''(X_i) \mathbf{z}))$

linear term is zero.

The results of this section will be employed in the remainder of this chapter to motivate and justify approaches to inference based on comparisons of model fits.

### 3.3 Comparing Models via Fitted Values

Section 3.2 defined seven models and gave expressions for the fitted values and their asymptotic biases under each model. Having described the estimation stage of the modelling process, questions of model inference now arise, i.e. how can comparisons be made which will enable us to select from these seven models the most appropriate one for a given set of data?

This section describes methods of inference via tests comparing two different model fits. 2 comparisons are considered: the difference in the residual sums of squares and a direct comparison of the fitted values. In a linear parametric setting these two approaches are equivalent as Section 3.3.1 shows. Section 3.3.2 describes these approaches in the context of nonparametric models, highlighting why the direct comparison of fitted values is preferable to the residual sums of squares approach.

#### 3.3.1 Comparing linear fits

Consider a linear model  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ , where  $\mathbf{y}$  is an  $n$  vector of responses,  $\mathbf{X}$  is an  $n \times p$  design matrix,  $\boldsymbol{\beta}$  is a  $p \times 1$  parameter matrix and  $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$ . A common test proceeds by defining some restriction on the values that the

(unknown) parameters can take, i.e.  $H_R : \mathbf{A}\beta = \mathbf{c}$ , where  $\mathbf{A}$  is a specified  $q \times p$  matrix and  $\mathbf{c}$  is a specified  $q$ -vector ( $q \leq p$ ). For example  $\mathbf{A} = \mathbf{I}_p$  and  $\mathbf{c} = [0, 0, \dots, 0]^T$  is equivalent to testing whether all the components of  $\beta$  are zero. We shall refer to the model with  $p$  parameters as the *full model* and the one with  $q$  variables as the *reduced model*.

The residual sum of squares of the full model is defined as

$$RSS_F = (\mathbf{y} - \mathbf{X}\hat{\beta}_F)^T(\mathbf{y} - \mathbf{X}\hat{\beta}_F)$$

where  $\hat{\beta}_F$  is the parameter estimate from the full model. It can be shown (Seber (1977) pp.85) that the parameter estimate of the reduced model can be defined as

$$\hat{\beta}_R = \hat{\beta}_F + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T[\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T]^{-1}(\mathbf{c} - \mathbf{A}\hat{\beta}_F)$$

and thus the difference in residual sum of squares of the two models can be written as

$$\begin{aligned} RSSD &= RSS_R - RSS_F \\ &= (\mathbf{y} - \mathbf{X}\hat{\beta}_R)^T(\mathbf{y} - \mathbf{X}\hat{\beta}_R) - (\mathbf{y} - \mathbf{X}\hat{\beta}_F)^T(\mathbf{y} - \mathbf{X}\hat{\beta}_F) \\ &= (\mathbf{A}\hat{\beta}_F - \mathbf{c})^T[\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T]^{-1}(\mathbf{A}\hat{\beta}_F - \mathbf{c}). \end{aligned}$$

This difference in residual sum of squares forms the backbone of methods of inference such as the F test for model comparisons.

Consider now a comparison based not on residuals but on a direct comparison of fitted values.

$$\begin{aligned} CFV &= (\mathbf{X}\hat{\beta}_F - \mathbf{X}\hat{\beta}_R)^T(\mathbf{X}\hat{\beta}_F - \mathbf{X}\hat{\beta}_R) \\ &= (\hat{\beta}_F - \hat{\beta}_R)^T\mathbf{X}^T\mathbf{X}(\hat{\beta}_F - \hat{\beta}_R) \\ &= [(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T[\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T]^{-1}(\mathbf{A}\hat{\beta}_F - \mathbf{c})]^T \\ &\quad \mathbf{X}^T\mathbf{X}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T[\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T]^{-1}(\mathbf{A}\hat{\beta}_F - \mathbf{c})] \\ &= (\mathbf{A}\hat{\beta}_F - \mathbf{c})^T[\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T]^{-1}\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &\quad \mathbf{A}^T[\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T]^{-1}(\mathbf{A}\hat{\beta}_F - \mathbf{c}) \end{aligned}$$

$$\begin{aligned}
&= (\mathbf{A}\hat{\beta}_F - \mathbf{c})^T [\mathbf{A}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{A}^T]^{-1} (\mathbf{A}\hat{\beta}_F - \mathbf{c}) \\
&= RSSD.
\end{aligned}$$

This demonstrates that the two approaches to model comparison, one based on residuals the other on fitted values, are equivalent when two linear model fits are related by a linear parameter constraint.

### 3.3.2 Nonparametric case

Consider a nonparametric additive model,  $Y_i = \alpha + \sum_{j=1}^p f_j(X_{ij}) + \varepsilon_i$ , where  $\alpha$  is a constant and the  $f_j$  are arbitrary smooth univariate functions, one for each predictor  $X_{.j}$  and  $\varepsilon_i \sim N(0, \sigma^2)$ .

As previously, we label this model with  $p$  terms the *full* model and consider a second *reduced* model, involving the first  $q$  ( $q < p$ ) predictors of the full model, i.e.  $Y_i = \alpha + \sum_{j=1}^q f_j(X_{ij}) + \varepsilon_i$ . To simplify notation let

$$f_F(\mathbf{x}_{i.}) = \alpha + f_1(X_{i1}) + f_2(X_{i2}) + \cdots + f_{(q-1)}(X_{i(q-1)}) + f_q(X_{iq}) + \cdots + f_p(X_{ip})$$

and

$$\hat{f}_F(\mathbf{x}_{i.}) = \hat{\alpha} + \hat{f}_1^F(X_{i1}) + \hat{f}_2^F(X_{i2}) + \cdots + \hat{f}_{(q-1)}^F(X_{i(q-1)}) + \hat{f}_q^F(X_{iq}) + \cdots + \hat{f}_p^F(X_{ip})$$

where  $\hat{f}_j^F$  is an estimate of  $f_j$  when the full model is fitted. Similarly, let

$$f_R(\mathbf{x}_{i.}) = \alpha + f_1(X_{i1}) + f_2(X_{i2}) + \cdots + f_q(X_{iq})$$

and

$$\hat{f}_R(\mathbf{x}_{i.}) = \hat{\alpha} + \hat{f}_1^R(X_{i1}) + \hat{f}_2^R(X_{i2}) + \cdots + \hat{f}_q^R(X_{iq})$$

where  $\hat{f}_j^R$  is an estimate of  $f_j$  when the reduced model is fitted.

Under the assumption that the reduced model is in fact correct, i.e.  $Y_i = f_R(\mathbf{x}_{i.}) + \varepsilon_i$ , we can define the bias of the two estimators  $\hat{f}_R(\mathbf{x}_{i.})$  and  $\hat{f}_F(\mathbf{x}_{i.})$  as

$$b_R(\mathbf{x}_{i.}) = e_R(X_{i.}) - f_R(\mathbf{x}_{i.})$$



and

$$b_F(\mathbf{x}_{i.}) = e_F(\mathbf{x}_{i.}) - \hat{f}_R(\mathbf{x}_{i.})$$

where  $e_R(\mathbf{x}_{i.})$  and  $e_F(\mathbf{x}_{i.})$  are the expectations with respect to the reduced model of the estimators  $\hat{f}_R(\mathbf{x}_{i.})$  and  $\hat{f}_F(\mathbf{x}_{i.})$  respectively. Also note that the estimators  $\hat{f}_R(\mathbf{x}_{i.})$  and  $\hat{f}_F(\mathbf{x}_{i.})$  can be written in terms of their expectations and a residual term. That is

$$\hat{f}_R(\mathbf{x}_{i.}) = e_R(\mathbf{x}_{i.}) + \hat{f}_R^\varepsilon(\mathbf{x}_{i.})$$

and

$$\hat{f}_F(\mathbf{x}_{i.}) = e_F(\mathbf{x}_{i.}) + \hat{f}_F^\varepsilon(\mathbf{x}_{i.}).$$

Using these decompositions the difference in residual sum of squares of the two models can be written as

$$\begin{aligned} RSSD &= \sum_{i=1}^n (f_R(\mathbf{x}_{i.}) + \varepsilon_{Ri} - e_R(\mathbf{x}_{i.}) - \hat{f}_R^\varepsilon(\mathbf{x}_{i.}))^2 - \\ &\quad \sum_{i=1}^n (f_R(\mathbf{x}_{i.}) + \varepsilon_{Ri} - e_F(\mathbf{x}_{i.}) - \hat{f}_F^\varepsilon(\mathbf{x}_{i.}))^2 \\ &= \sum_{i=1}^n (\varepsilon_{Ri} - \hat{f}_R^\varepsilon(\mathbf{x}_{i.}) - b_R(\mathbf{x}_{i.}))^2 - \\ &\quad \sum_{i=1}^n (\varepsilon_{Ri} - \hat{f}_F^\varepsilon(\mathbf{x}_{i.}) - b_F(\mathbf{x}_{i.}))^2 \end{aligned}$$

Using the same notation we can define a comparison of the fitted values as

$$\begin{aligned} CFV &= \sum_{i=1}^n (\hat{f}_R(\mathbf{x}_{i.}) - \hat{f}_F(\mathbf{x}_{i.}))^2 \\ &= \sum_{i=1}^n (e_R(\mathbf{x}_{i.}) + \hat{f}_R^\varepsilon(\mathbf{x}_{i.}) - e_F(\mathbf{x}_{i.}) - \hat{f}_F^\varepsilon(\mathbf{x}_{i.}))^2 \\ &= \sum_{i=1}^n (b_R(\mathbf{x}_{i.}) - b_F(\mathbf{x}_{i.}) + \hat{f}_R^\varepsilon(\mathbf{x}_{i.}) - \hat{f}_F^\varepsilon(\mathbf{x}_{i.}))^2 \end{aligned}$$

Comparing these expressions for RSSD and CFV we see that if the bias terms  $b_R$  and  $b_F$  were equal they would cancel directly in CFV since they do not appear combined with other terms as in the cross-product term of RSSD.

Returning to the bivariate class of models, Section 3.2.8 showed that bias is present in the fitted values of models involving nonparametric components. Table 3.3 also demonstrated that under certain conditions the asymptotic bias

of the 'reduced' and 'full' models were equivalent when the reduced model was in fact true. These results suggest that the use of model comparisons based on CFV may be more suited to handling the bias inherent in nonparametric fits. The next section examines how the CFV statistic could be used to define a method of inference.

### 3.4 Distributional Properties of CFV

The last section demonstrated how the properties of the CFV statistic suggest its use for model comparisons using nonparametric model fits. The next step in defining methods of inference using CFV is to gain some insight into its distributional properties. Matrix notation, like that introduced in Section 3.2 once again will prove useful. In this discussion it is assumed that the covariates  $X$  and  $Z$  are independent. The case of dependence is considered via simulations in Section 4.3.3.

Consider one of the comparison of model fits listed in Table 3.2, where the model fits are given by:

$$\hat{m}_F(\mathbf{x}, \mathbf{z}) = \mathbf{S}_F \mathbf{y}$$

and

$$\hat{m}_R(\mathbf{x}, \mathbf{z}) = \mathbf{S}_R \mathbf{y}$$

where  $\mathbf{S}_F$  and  $\mathbf{S}_R$  are smoothing matrices which return the full and reduced model fits respectively.

Using this notation,

$$\begin{aligned} CFV &= \sum_{i=1}^n (\hat{m}_R(X_i, Z_i) - \hat{m}_F(X_i, Z_i))^2 \\ &= \mathbf{y}^T (\mathbf{S}_R - \mathbf{S}_F)^T (\mathbf{S}_R - \mathbf{S}_F) \mathbf{y} \\ &= \mathbf{y}^T \mathbf{A} \mathbf{y} \end{aligned} \tag{3.5}$$

where  $\mathbf{A} = (\mathbf{S}_R - \mathbf{S}_F)^T (\mathbf{S}_R - \mathbf{S}_F)$ .

Assuming that the reduced model is a true reflection of the form of the regression function, i.e.  $y = m_R(\mathbf{x}, \mathbf{z}) + \varepsilon$ ,

$$\begin{aligned} CFV &= (m_R(\mathbf{x}, \mathbf{z}) + \varepsilon)^T (\mathbf{S}_R - \mathbf{S}_F)^T (\mathbf{S}_R - \mathbf{S}_F) (m_R(\mathbf{x}, \mathbf{z}) + \varepsilon) \\ &= m_R(\mathbf{x}, \mathbf{z})^T (\mathbf{S}_R - \mathbf{S}_F)^T (\mathbf{S}_R - \mathbf{S}_F) m_R(\mathbf{x}, \mathbf{z}) + \varepsilon^T (\mathbf{S}_R - \mathbf{S}_F)^T (\mathbf{S}_R - \mathbf{S}_F) \varepsilon. \end{aligned}$$

When a linear smoother is applied to the true underlying function's values at the design points it returns the expected values of the fitted model. Combining this with the property that the bias of the two estimates is equal when the reduced model is true gives:

$$m_R(\mathbf{x}, \mathbf{z})^T (\mathbf{S}_R - \mathbf{S}_F)^T (\mathbf{S}_R - \mathbf{S}_F) m_R(\mathbf{x}, \mathbf{z}) = 0$$

and therefore

$$CFV = \varepsilon^T (\mathbf{S}_R - \mathbf{S}_F)^T (\mathbf{S}_R - \mathbf{S}_F) \varepsilon = \varepsilon^T \mathbf{A} \varepsilon. \quad (3.6)$$

This demonstrates that a comparison of fitted values for each of the comparisons listed in Table 3.2 reduces to a quadratic form in zero mean random variables ( $\varepsilon$ ) when the reduced form of the regression function is true and the biases of the two fitted models (reduced and full) are equal. This result is attractive from an inferential viewpoint since the distributions of quadratic forms are well understood.

Chapter 29 of Johnson & Kotz (1972) gives a detailed description of the distributional properties of a quadratic form in normal variates. For the case here of centred (zero mean) random variables  $\varepsilon$ , the key result is the expression for the cumulants, namely

$$\kappa_s = 2^{s-1} (s-1)! \text{tr}\{(\mathbf{V}\mathbf{A})^j\}. \quad (3.7)$$

where  $\text{tr}\{\cdot\}$  denotes the trace operator, and  $\mathbf{V}$  is the covariance matrix of the normal variate. Since our quadratic form has reduced to  $\varepsilon^T \mathbf{A} \varepsilon$ ,  $\mathbf{V} = \sigma^2 \mathbf{I}$ .

Through these expressions for the cumulants we have instant access to properties such as the mean ( $\kappa_1$ ) and variance ( $\kappa_2$ ) of the distribution. Thus these quantities (and higher order cumulants) can be matched with a more accessible distribution in order to provide convenient calculations of p-values of observed test statistics. A shifted and scaled  $\chi^2$  distribution is suggested by authors such as Bowman and Azzalini (1997) (see Section 3.5.3).

### 3.5 Tests Under Investigation

Having outlined several comparisons of fitted models the next step is to devise methods of assessing the significance of observed differences. Both the RSSD and CFV statistics can be expressed in matrix notation as  $\mathbf{y}^T \mathbf{A} \mathbf{y}$ , and the last Section showed that CFV reduces to  $\boldsymbol{\varepsilon}^T \mathbf{A} \boldsymbol{\varepsilon}$  under certain conditions. Since  $\boldsymbol{\varepsilon}$  comprises i.i.d. normal random variable with constant variance  $\sigma^2$ , clearly this unknown quantity affects the distributional properties of both comparisons of model fits.

Chapter 2 described a number of estimators of  $\sigma^2$ . These ranged from an adjusted residual sum of squares (analogous to linear regression) to estimators which use pseudoresiduals defined by ‘differencing’ the response. Each of these estimators, however, could be written in the matrix form  $\mathbf{y}^T \mathbf{B} \mathbf{y}$  where  $\mathbf{B}$  is an  $n \times n$  matrix of constants depending on the design space. The obvious test statistic which attempts to scale out the unknown variance from the model comparison statistic is therefore of the form

$$\frac{\text{comparison of model fit statistic}}{\hat{\sigma}^2} = \frac{\mathbf{y}^T \mathbf{A} \mathbf{y}}{\mathbf{y}^T \mathbf{B} \mathbf{y}}.$$

A number of reference distributions have been proposed for tests of this type. An approximate F-test has been suggested by analogy with the linear parametric setting. A two-moment correction which attempts to address the inaccuracy of the F-approximation has also been suggested (Hastie and Tibshirani 1990). Thirdly, distributional results are available directly from the theory of quadratic forms of normal random variables. Sections 3.5.1-3.5.3 consider each of these approaches, defining the reference distributions to be used in each case to assess observed values of the test statistics.

### 3.5.1 Approximate F-test

One approach to model comparisons analogous to the F test of linear modelling has been suggested by authors such as Hastie and Tibshirani (1990). It requires a test statistic of the form,

$$F = \frac{\mathbf{y}^T \mathbf{A} \mathbf{y}}{\text{tr}(\mathbf{A})} / \frac{\mathbf{y}^T \mathbf{B} \mathbf{y}}{\text{tr}(\mathbf{B})}, \quad (3.8)$$

whose denominator reflects the RSS based estimators of error variance employed frequently in linear regression<sup>6</sup>.

Under the assumption that the reduced model is true, a natural reference distribution for 3.8 is

$$F_{\text{tr}(\mathbf{A}), \text{tr}(\mathbf{B})}. \quad (3.9)$$

This approach can be understood mostly clearly in the context of RSS comparisons, although it applies equally when  $\mathbf{A}$  represents a CFV comparisons. Let

$$\begin{aligned} F_{RSS} &= \frac{\mathbf{y}^T (\mathbf{S}_R - \mathbf{S}_F)^T (\mathbf{S}_R - \mathbf{S}_F) \mathbf{y}}{\text{tr}((\mathbf{S}_R - \mathbf{S}_F)^T (\mathbf{S}_R - \mathbf{S}_F))} / \frac{\mathbf{y}^T (\mathbf{I} - \mathbf{S}_F)^T (\mathbf{I} - \mathbf{S}_F) \mathbf{y}}{\text{tr}((\mathbf{I} - \mathbf{S}_F)^T (\mathbf{I} - \mathbf{S}_F))} \\ &= \frac{RSS_R - RSS_F}{df_R - df_F} / \frac{RSS_F}{df_F} \end{aligned} \quad (3.10)$$

where  $RSS_R$  and  $RSS_F$  are the residual sum of squares under the reduced and full model respectively (as defined earlier) and  $df_R$  and  $df_F$  are approximated error degrees of freedom of the two fits (defined in Section 2.2). The observed value is then compared with the  $F_{df_R - df_F, df_F}$  distribution to obtain a p-value for the test that the reduced model is adequate.

This approach naturally employs  $\text{tr}((\mathbf{I} - \mathbf{S})^T (\mathbf{I} - \mathbf{S})) = n - \text{tr}(2\mathbf{S} - \mathbf{S}^T \mathbf{S})$  as the definition of the error degrees of freedom. In Section 2.2 we saw that

$$E \left( \frac{RSS_F}{\sigma^2} \right) \approx \frac{1}{\sigma^2} E(\boldsymbol{\varepsilon}^T \mathbf{B}_F \boldsymbol{\varepsilon})$$

---

<sup>6</sup>The notation F for the observed value reflects its reference distribution

$$\begin{aligned}
&= \text{tr}\{\mathbf{B}_F\} \\
&= \text{tr}\{(I - \mathbf{S}_F)^T(I - \mathbf{S}_F)\} \\
&= (n - \text{tr}\{2\mathbf{S}_F - \mathbf{S}_F^T\mathbf{S}_F\}) \\
&= df_F.
\end{aligned}$$

and similarly

$$\begin{aligned}
E\left(\frac{RSS_R - RSS_F}{\sigma^2}\right) &\approx \frac{1}{\sigma^2}E(\boldsymbol{\varepsilon}^T(\mathbf{B}_R - \mathbf{B}_F)\boldsymbol{\varepsilon}) \\
&= \text{tr}\{\mathbf{B}_R - \mathbf{B}_F\} \\
&= \text{tr}\{2\mathbf{S}_F - \mathbf{S}_F^T\mathbf{S}_F\} - \text{tr}\{2\mathbf{S}_R - \mathbf{S}_R^T\mathbf{S}_R\} \\
&= df_R - df_F.
\end{aligned}$$

It is clear from these expressions where the approximations

$$\frac{RSS_F}{\sigma^2} \sim \chi_{\text{tr}\{\mathbf{B}_F\}}^2 \text{ and } \frac{RSS_R - RSS_F}{\sigma^2} \sim \chi_{\text{tr}\{\mathbf{B}_R - \mathbf{B}_F\}}^2$$

come from. These approximations account for the use of  $F_{\text{tr}\{\mathbf{B}_R - \mathbf{B}_F\}, \text{tr}\{\mathbf{B}_R\}}$  to assess the observed values of  $F_{RSS}$ . They also suggest distributions that can be used to assess model differences when  $\sigma^2$  is known.

When using difference based estimators of  $\sigma^2$ , however, this reference distribution is not appropriate. Recall Section 2.5 defined a difference based estimator in the form,  $\mathbf{y}^T\mathbf{B}\mathbf{y}$ , where estimating matrix  $\mathbf{B}$  is such that  $\text{tr}\{\mathbf{B}\} = 1$ . Although this form doesn't affect the calculation of an observed value according to Equation 3.8, the corresponding  $F$  reference distribution cannot be calculated since its second parameter will always be 1. That is, different difference based estimators (eg. a weighted and an unweighted version) will each yield different observed values but the same reference distribution (since  $\text{tr}\{\mathbf{B}\} = 1$  in both cases). Therefore the approximate  $F$  test is only applicable when  $RSS$  based estimators of  $\sigma^2$  are employed.

It should be stressed, however, that despite the frequent adoption and availability of this method, these results are not based on theoretical distributional

results, but rather on an analogy with approaches in linear regression. As such, their performance can be improved upon (see Section 3.5.2) and the performance will be scrutinised in Chapter 4.

### 3.5.2 Two-moment corrected F-test

A *two-moment corrected* F distribution can be used to improve the approximation of the F distribution to the true distribution of the test statistic (3.8). The corrected F distribution is

$$F \sim F_{\frac{tr(\mathbf{A})^2}{tr(\mathbf{A}^2)}, \frac{tr(\mathbf{B})^2}{tr(\mathbf{B}^2)}}. \quad (3.11)$$

This approach was originally motivated by the RSS approach (Cleveland and Devlin, 1989). Although the residual sum of squares will always contain bias in a nonparametric setting (a consequence of which is that exact distributional results are not available), the approximate  $F$  test above can be improved upon by applying a two moment correction to both the denominator and numerator of  $F_{RSS}$ . Consider the denominator where the second moment correction proceeds by finding a multiple of  $RSS_F$ , say  $k$ , such that the mean and variance of  $k.RSS_F$  matches that of a reference  $\chi^2$  distribution. One property of the  $\chi^2$  distribution is that its mean and variance are in the ratio 2:1. Therefore,  $k$  must satisfy the following:

$$\text{var} \left( k \cdot \frac{RSS_F}{\sigma^2} \right) = 2.E \left( k \cdot \frac{RSS_F}{\sigma^2} \right).$$

Since

$$\begin{aligned} \text{var} \left( \frac{RSS_F}{\sigma^2} \right) &\approx \frac{1}{\sigma^4} \text{var}(\epsilon^T \mathbf{B}_F \epsilon) \\ &= 2tr\{\mathbf{B}_F^2\}, \end{aligned}$$

$k$  is found by solving

$$2k^2 tr\{\mathbf{B}_F^2\} = 2k tr\{\mathbf{B}_F\}$$

to yield the two moment correcting factor

$$k = \frac{\text{tr}\{\mathbf{B}_F\}}{\text{tr}\{\mathbf{B}_F^2\}}.$$

The two moment approximation can therefore be written <sup>7</sup> as

$$\frac{\text{tr}\{\mathbf{B}_F\}}{\text{tr}\{\mathbf{B}_F^2\}} \frac{RSS_F}{\sigma^2} \sim \chi^2_{\frac{\text{tr}\{\mathbf{B}_F\}^2}{\text{tr}\{\mathbf{B}_F^2\}}}$$

Similarly, the two moment approximation for the numerator of  $F_{RSS}$  can be shown to be

$$\frac{\text{tr}\{\mathbf{B}_R - \mathbf{B}_F\}}{\text{tr}\{(\mathbf{B}_R - \mathbf{B}_F)^2\}} \frac{RSS_R - RSS_F}{\sigma^2} \sim \chi^2_{\frac{\text{tr}\{\mathbf{B}_R - \mathbf{B}_F\}^2}{\text{tr}\{(\mathbf{B}_R - \mathbf{B}_F)^2\}}}$$

Using these two distributional approximations, a *two-moment corrected*  $F$  statistic is defined with a numerator given by

$$\frac{\text{tr}\{\mathbf{B}_R - \mathbf{B}_F\}}{\text{tr}\{(\mathbf{B}_R - \mathbf{B}_F)^2\}} \frac{RSS_R - RSS_F}{\sigma^2} / \frac{\text{tr}\{\mathbf{B}_R - \mathbf{B}_F\}^2}{\text{tr}\{(\mathbf{B}_R - \mathbf{B}_F)^2\}} = \frac{RSS_R - RSS_F}{\sigma^2 \text{tr}\{\mathbf{B}_R - \mathbf{B}_F\}}$$

and a denominator similarly defined as

$$\frac{\text{tr}\{\mathbf{B}_F\}}{\text{tr}\{\mathbf{B}_F^2\}} \frac{RSS_F}{\sigma^2} / \frac{\text{tr}\{\mathbf{B}_F\}^2}{\text{tr}\{\mathbf{B}_F^2\}} = \frac{RSS_F}{\sigma^2 \text{tr}\{\mathbf{B}_F\}}.$$

Clearly the ratio of these two quantities will yield a test statistic identical to  $F_{RSS}$ . The difference here is that we use  $F_{\frac{\text{tr}\{\mathbf{B}_R - \mathbf{B}_F\}^2}{\text{tr}\{(\mathbf{B}_R - \mathbf{B}_F)^2\}}, \frac{\text{tr}\{\mathbf{B}_F\}^2}{\text{tr}\{\mathbf{B}_F^2\}}}$  as the reference distribution. These results also show that a  $\chi^2_{\frac{\text{tr}\{\mathbf{B}_F\}^2}{\text{tr}\{\mathbf{B}_F^2\}}}$  distribution could be used if  $\sigma^2$  was known.

It was noted above, in the context of the approximate  $F$  test, that difference based estimators of  $\sigma^2$  are not of a form which naturally fits with a reference  $F$  distribution. In the case of a two moment correction, however, this is not

---

<sup>7</sup>Note: this differs from an equivalent expression in Equation 3.30 of Hastie and Tibshirani (1990) since the degrees of freedom of the reference distribution are incorrect in Hastie and Tibshirani (1990)



a difficulty. Consider two *different* estimating matrices,  $\mathbf{B}$  and  $\mathbf{B}'$ , such that  $\mathbf{B} = \mathbf{B}'/\text{tr}(\mathbf{B}')$  ( $\text{tr}(\mathbf{B}) = 1$  and  $\text{tr}(\mathbf{B}') \neq 1$ ), yielding the *same* estimate  $\hat{\sigma}^2 = \mathbf{y}^T \mathbf{B} \mathbf{y} = \mathbf{y}^T \mathbf{B}' \mathbf{y} / (\text{tr}(\mathbf{B}'))$ . The corrected  $F$  test assumes that the estimate of  $\sigma^2$  is of the form  $\mathbf{y}^T \mathbf{B}' \mathbf{y} / \text{tr}(\mathbf{B}')$  although the difference based estimator is of the form  $\mathbf{y}^T \mathbf{B} \mathbf{y}$ . But consider the degrees of freedom employed in the corrected  $F$  distribution,

$$\frac{\text{tr}(\mathbf{B}')^2}{\text{tr}(\mathbf{B}'^2)} = \frac{\text{tr}(\mathbf{B}')^2 / \text{tr}(\mathbf{B}')^2}{\text{tr}(\mathbf{B}'^2) / \text{tr}(\mathbf{B}')^2} = \frac{\text{tr}(\mathbf{B})^2}{\text{tr}(\mathbf{B}^2)}.$$

Therefore, we can calculate the degrees of freedom, defined in terms of  $\mathbf{B}'$ , from the matrix  $\mathbf{B}$  which is available for the difference based estimator of  $\sigma^2$ . Therefore, the corrected  $F$  distribution is applicable to both forms of the comparisons of model fits and both forms of the estimator of  $\sigma^2$ . Its performance will be investigated via a simulation study in Sections 4.2-4.3.

### 3.5.3 Quadratic form approach

The third and final approach to assessing the significance of observed differences uses the observed value of:

$$Q = \frac{\mathbf{y}^T \mathbf{A} \mathbf{y}}{\mathbf{y}^T \mathbf{B} \mathbf{y}}. \quad (3.12)$$

The reference distribution is derived from the theory of quadratic forms in standard normal random variables, introduced in Section 3.4. As such, model comparisons using CFV are the motivation for such an approach, although the distributional calculations can be performed on differences of RSS's and with any  $\hat{\sigma}^2$  under consideration.

To aid the derivation of the reference distribution, rearrange Equation 3.12 to become

$$\mathbf{y}^T \mathbf{C} \mathbf{y} = 0 \quad \text{where} \quad \mathbf{C} = \mathbf{A} - Q\mathbf{B}$$

If the bias terms in  $\mathbf{y}^T \mathbf{A} \mathbf{y}$  cancel and the denominator itself is an unbiased estimator of  $\sigma^2$ , then we can write

$$\mathbf{y}^T \mathbf{C} \mathbf{y} \approx \boldsymbol{\epsilon}^T \mathbf{C} \boldsymbol{\epsilon} \quad \text{where} \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}),$$

that is, the test statistic is approximately a quadratic form in zero mean, unit variance normal random variables.

We can, therefore, calculate a p-value based on an observed value  $Q$  using the properties of a quadratic form  $\boldsymbol{\epsilon}^T \mathbf{C} \boldsymbol{\epsilon}$ . Rather than calculate an exact p-value however, for the purposes of hypothesis testing, p-values can be calculated by assigning the first three or four moments of the distribution of  $\mathbf{y}^T \mathbf{C} \mathbf{y}$  to a more convenient distribution. In particular Bowman & Azzalini (1997, pp. 103) describe the use of a shifted and scaled  $\chi^2$  distribution,  $a\chi_b^2 + c$  say, for this purpose.

Equation 3.7 defines the cumulants of a centred (zero mean) quadratic form. In particular the first three (approximate) cumulants of  $Q$  are given by

$$\begin{aligned} \kappa_1 &= \text{tr}(\mathbf{C}) \\ \kappa_2 &= 2\text{tr}(\mathbf{C}^2) \\ \kappa_3 &= 8\text{tr}(\mathbf{C}^3) \end{aligned}$$

Using these expressions and the fact that for this test statistic  $\mathbf{V} = \mathbf{I}$  (independent error with unit variance) we can match the first three moments of the two distributions by setting

$$a = |\kappa_3|/(4\kappa_2), \quad b = (8\kappa_2^3)\kappa_3^3, \quad c = \kappa_1 - ab.$$

The p-value of the observed test statistic,  $Q$ , can then be approximated as  $1 - q$ , where  $q$  is the probability lying below the point  $-c/a$  in a  $\chi^2$  distribution with  $b$  degrees of freedom.

**Table 3.5.** Components of tests to compare model fits

CMF	Ref. Dist.	$\hat{\sigma}^2$
RSS	approx. F	RSS based
		RSS based - undersmoothed
	corr. F	RSS based
		RSS based - undersmoothed
		difference based
	Quadratic Form	RSS based
		RSS based - undersmoothed
		difference based
CFV	approx. F	RSS based
		RSS based - undersmoothed
	corr. F	RSS based
		RSS based - undersmoothed
		difference based
	Quadratic Form	RSS based
		RSS based - undersmoothed
		difference based

### 3.5.4 Summary

In summary then, there are three stages in the process of model inference via the comparisons of model fits:

1. calculate a statistic which compares the two model fits (CMF)
2. divide by an estimate of the error variance ( $\hat{\sigma}^2$ )
3. assess the resulting test statistic via a reference distribution (Ref. Dist.)

At each stage there are a number of options available, as this chapter has described. Table 3.5 summarises the combinations of choices within these three stages each of which defines a different approach to model inference. These methods are investigated in Chapter 4 via a simulation study.

## Chapter 4

# Inference Amongst a Class of Bivariate Nonparametric Models: Simulation Results

### 4.1 Introduction

The simulation studies described in this chapter consider each of the model comparisons listed in Table 3.2. Section 2.7.2 defined the regression functions underlying simulated data used to investigate estimators of  $\sigma^2$ . These same regression functions will be used here to simulate data to assess both the size and the power performance of each test. Appendix A describes the principal S-Plus functions created and used throughout the simulation studies.

The size of each test is assessed by generating sets of data from the *reduced model* and comparing the fit of this true model class with a model fit belonging to the *full model* class. Power simulations are performed by generating data from the *full model* and comparing a fit of this true form with one corresponding to the *reduced model* class. The true reference distribution of the test statistics will yield p-values distributed  $U(0,1)$  under the null hypothesis. More relevant to hypothesis testing is the lower tail of the distribution of the test statistic. Therefore, the proportion of the simulated p-values less than 0.05, the *empirical size*, are used to assess the performance of the different tests over a range of

conditions.

As an illustration, Figure 4.1 displays histograms of p-values from one of the simulations described later<sup>1</sup>. Each panel represent a different test, i.e. a different combination of a comparison of model fits, an estimator of  $\sigma^2$  and a reference distribution. The top left panel shows results obtained using the standard  $F$  test approach. Clearly there is substantial skewness resulting in a conservative test. The top right panel shows that a two-moment corrected F distribution does not offer a significant improvement. The bottom panels, however, show the marked effect of using a difference based estimator of  $\sigma^2$  and the effect of using the CFV test statistic. To assess the empirical size of each test for substantial departures from the specified  $\alpha = 0.05$  level, a 99% confidence interval based on a binomial distribution with  $n$  =(number of simulations) and  $p = 0.05$  is used. The limits of these intervals are shown by the horizontal lines in Figure 4.1 and clearly demonstrate that the tests of the top two panels are inconsistent at the  $\alpha = 0.05$  level whereas the bottom two test approaches return consistent results.

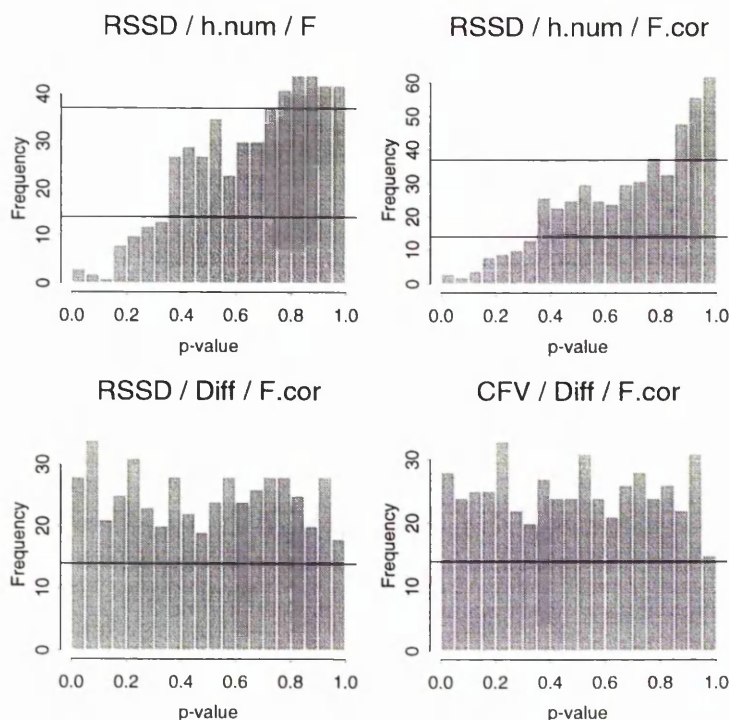
In the following sections, each test will be considered over a range of settings achieved by varying the following ‘settings’:

- ◇ design space
- ◇ sample size
- ◇ error variance
- ◇ smoothing parameter.

Furthermore, the two scenarios of  $\sigma^2$  known and  $\sigma^2$  estimated are considered separately. Although the former is rarely encountered in practice, it is useful to consider this case since it highlights the behaviour of the comparison of model fits statistics. The two broad classes of fixed-regular and random designs are considered separately in Sections 4.2 & 4.3 respectively.

---

<sup>1</sup>These results are ‘size’ results for a 1d.sm vs. 2d.am model comparison using a regular grid,  $n = 100$ ,  $\sigma^2 = 0.0025$  and  $h = 0.1$ , taken from Section 4.2.2. They are presented only for illustrative purposes only. A detailed exploration of the different approaches under different conditions is the aim of the simulation studies.



**Figure 4.1.** P-values obtained from the simulation study described in Section 4.2.2 showing the effect of the different choices of Comparison of Model Fits/Estimator of  $\sigma^2$ /Reference Distribution on the p-values generated over 500 simulations. The horizontal lines indicate a 99% confidence interval for the frequencies if the distribution of p-values was Uniform

## 4.2 Model Comparisons Simulation Study: Regular Grid Design

### 4.2.1 Regular grid: $\sigma^2$ known

This section will investigate model comparisons using the *known* value of  $\sigma^2$ , thereby focusing on the finite sample behaviour of RSSD and CFV statistics to assess the differences in the fits.

**Table 4.1.** Summary of the settings and the test procedures used in the simulations of Section 4.2.1.

Data conditions	
design space	regular grid over $[0, 1]^2$
sample size	49 design points
	100 design points
error distribution	Normal (zero mean)
error variance	0.01
	0.04
number of simulations	1000
Test approaches	
model comparisons	2d.am vs. 2d.sm
	1d.sm vs. 2d.am
	1d.sm vs. semi.par.x1
	semi.par.x1 vs. 2d.am
	st.line vs. semi.par.x2
smoothing parameter	0.05 & 0.15
variance estimator ( $\sigma^2$ )	NONE (true value used)
comparisons of model fits	RSSD & CFV
reference distribution	QF, $\chi^2$ , corrected $\chi^2$

### Investigating size

The empirical sizes based on 1000 simulations, using the tests and conditions of Table 4.1, are listed in Table B.1 (see Appendix B) and summarised graphically in Figure 4.2. If the test statistic and the reference distribution were well matched then under the null hypothesis (reduced model) the empirical sizes ( $\times 1000$ ) would be distributed  $\text{Bin}(1000, 0.05)$ . Figure 4.2 indicates the corresponding 99% confidence interval, (0.035, 0.067), by horizontal solid lines. This aids in the assessment of the difference between the observed sizes and the expected sizes of 0.05.

Using Figure 4.2, the following properties are observed in the regular grid setting with knowledge of  $\sigma^2$ :

**RSSD vs. CFV:** this comparison, at the heart of the simulation study, shows

that there is little to distinguish between the performance of the CFV statistics and the RSSD statistics.

**reference distributions:** both the corrected  $\chi^2$  distribution and that derived from the quadratic form formulation return consistent results. The standard  $\chi^2$  distribution fails to describe both test statistics for the model comparisons when the underlying model has effects in both  $X$  and  $Z$ .

**smoothing parameter** the two levels of smoothing parameter do not return markedly different results.

### Investigating power

Section 4.2.1 examined the performance of the size of the six combinations of two test statistics (RSSD and CFV) and three reference distributions (F, F.corr and QF) using the true value of  $\sigma^2$  and data simulated over a regular grid. This section completes the investigation for this case by considering the power performance, i.e. the ability of the tests to detect departures from the null model.

The power of these tests is related, amongst other things<sup>2</sup>, to the nature of the regression function's departure from the null model. Extreme cases are when the true regression function is only a slight deviation from the null model, in which case any test is unlikely to detect the difference (power=0), and, conversely, when the regression function is substantially different from the null model the tests will always detect a difference (power=1). It is necessary, therefore, to simulate data which will return empirical power away from these two extremes. The following regression functions were found to yield such results when they were used to define the full models listed in Table 3.2.

**semi.par.x1**  $m(X, Z) = 1 + X - 0.75 \exp(-0.5((X - 0.5)^2)/0.01) + \frac{1}{4}Z$

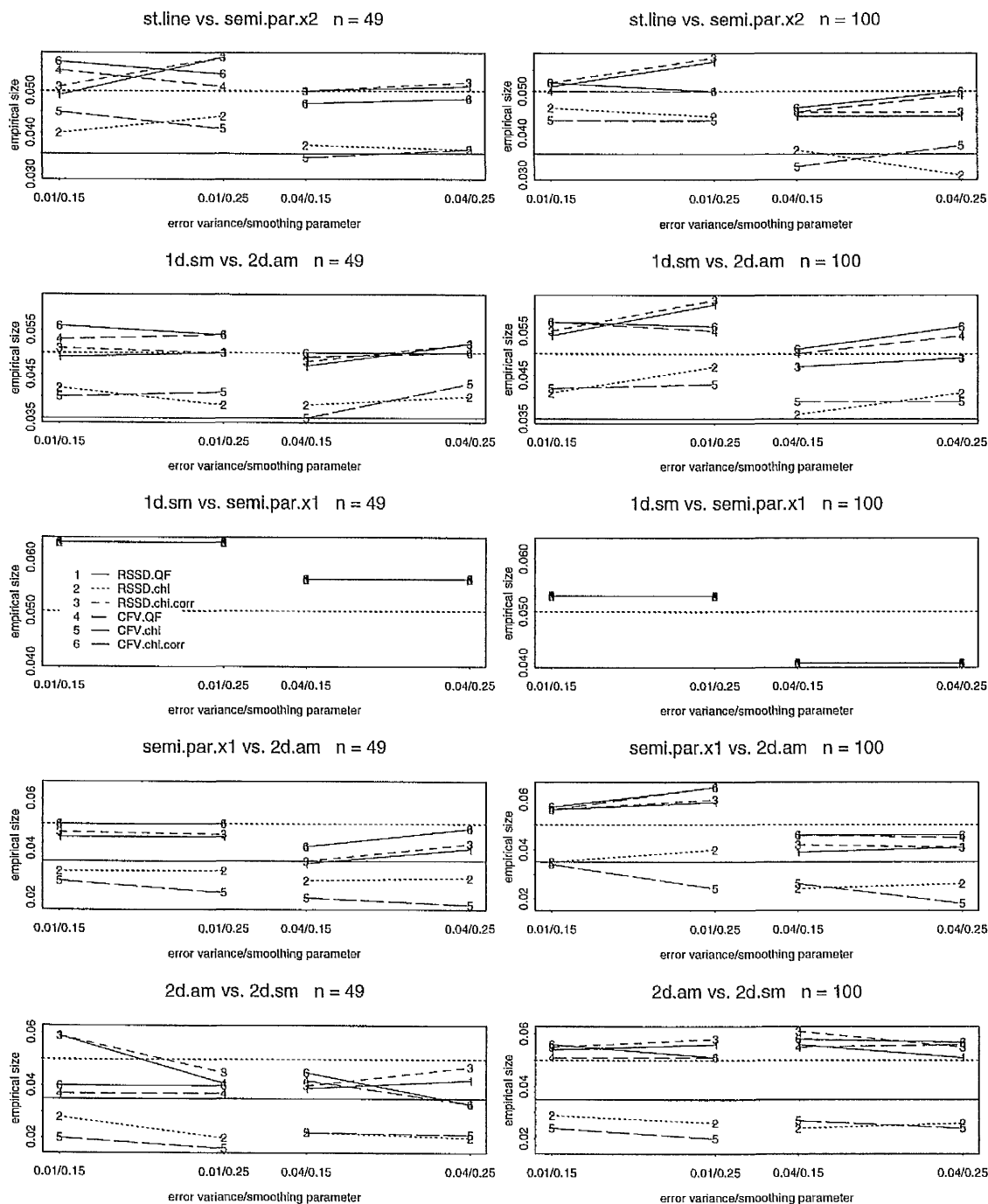
**semi.par.x2**  $m(X, Z) = \frac{1}{6}(1 + X - 0.75 \exp(-0.5((X - 0.5)^2)/0.01)) + Z$

**2d.am**  $m(X, Z) = 1 + X - 0.75 \exp(-0.5((X - 0.5)^2)/0.01) +$   
 $3.5cm \frac{1}{4}(1 + Z + 0.75 \exp(-0.5((Z - 0.5)^2)/0.01))$

---

<sup>2</sup>Another important factor is the error variance of the model.





**Figure 4.2.** Empirical sizes of tests of model comparisons using known  $\sigma^2$ . Design points form a regular square grid. The results listed are the proportion of 1000 simulated p-values less than 0.05 under the reduced model. The solid horizontal lines (when shown) indicate a 99% confidence interval from  $\text{Bin}(1000, 0.05)$ .

**2d.sm**  $m(X, Z) = \frac{1}{20}(-X + \Psi(X, Z, 0.3, 0.3, 0.1, 0.1, 0) + Z -$   
**3.5cm**  $\Psi(X, Z, 0.7, 0.7, 0.1, 0.1, 0))$

where  $\Psi(\cdot)$  is the probability density function of a bivariate normal distribution,

$$\Psi(X, Z, \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{e^{\frac{-1}{2(1-\rho^2)} \left( \left( \frac{X-\mu_1}{\sigma_1} \right)^2 - 2\rho \frac{X-\mu_1}{\sigma_1} \frac{Z-\mu_2}{\sigma_2} + \left( \frac{Z-\mu_2}{\sigma_2} \right)^2 \right)}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}.$$

The results listed in Table B.2 are summarised in Figure 4.3 and are based on 500 sets of data simulated over regular grids. The values themselves are quite arbitrary, since they mostly reflect the degree of departure from the null model. In summary, the results reveal the following:

**RSSD vs. CFV:** there is very little to distinguish between the different tests' power performances. This shows that the test statistics based on comparisons of fitted values do not suffer poorer power than their residual sum of squares based counterparts.

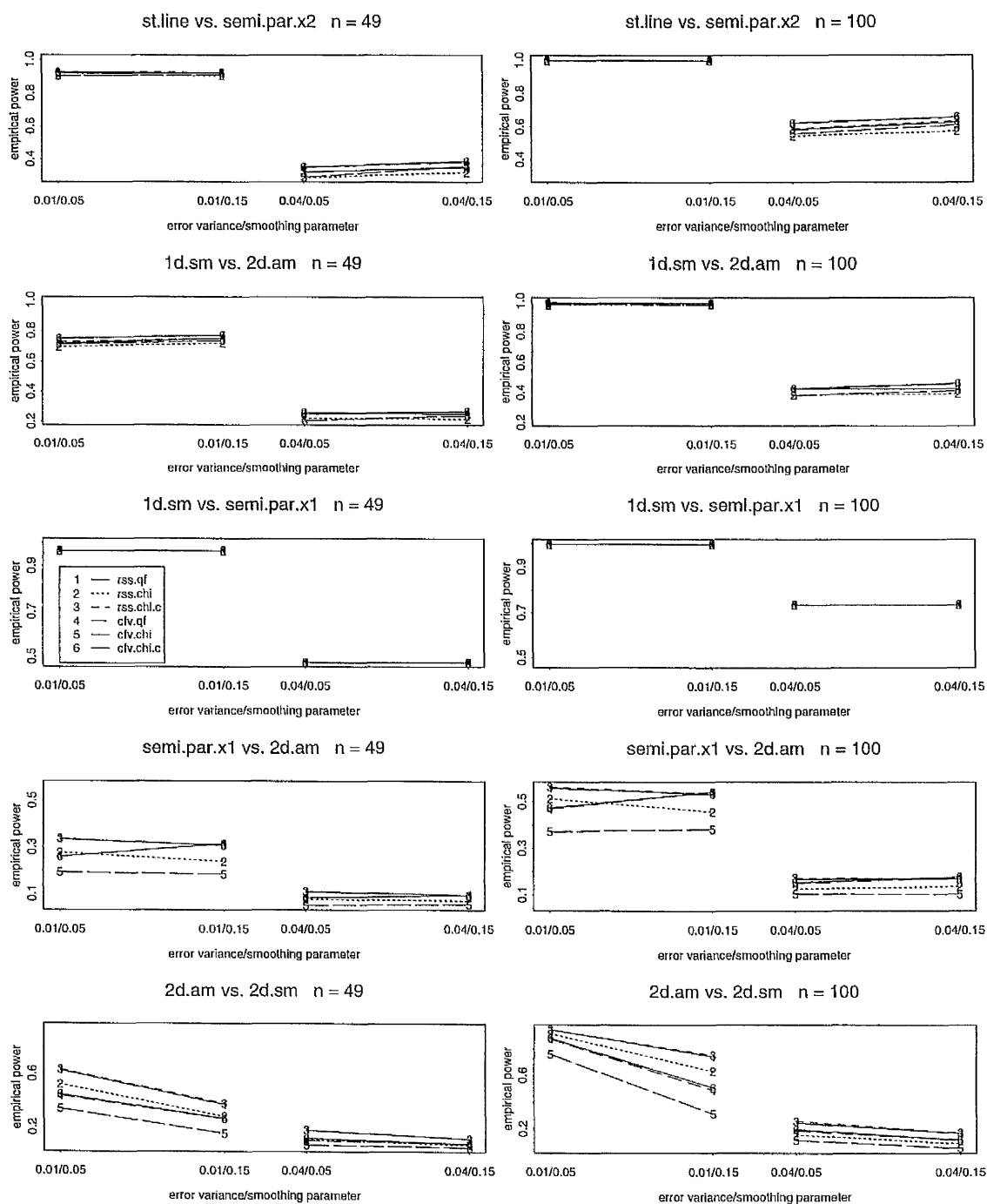
**reference distributions:** there is little to distinguish between the corrected  $\chi^2$  test and the quadratic form related distribution. In some cases, the  $\chi^2$  reference distribution yields lower values of power.

**smoothing parameter:** there is little evidence of an effect of increasing the smoothing parameter level with the exception being when the underlying regression surface is a bivariate smooth. In this case the effect of increasing  $h$  is to decrease the empirical power.

### 4.2.2 Regular grid: $\sigma^2$ estimated

When  $\sigma^2$  is unknown, as is the case nearly always in practice, an estimator of  $\sigma^2$  must also be incorporated into the test procedure. Hence there are three (four with smoothing parameter) choices concerning the comparison of different model fits, as described in Section 3.5.4 and summarised in Table 3.5.

Table 4.2 summarises the simulated data conditions and the test approaches reported in this section.



**Figure 4.3.** Power results of tests for model comparisons using known  $\sigma^2$ . Design points form a regular square grid. The results listed are the proportion of 500 simulated p-values less than 0.05 under the full model.

**Table 4.2.** Summary of the settings and the test procedures used in the simulations of Section 4.2.2.

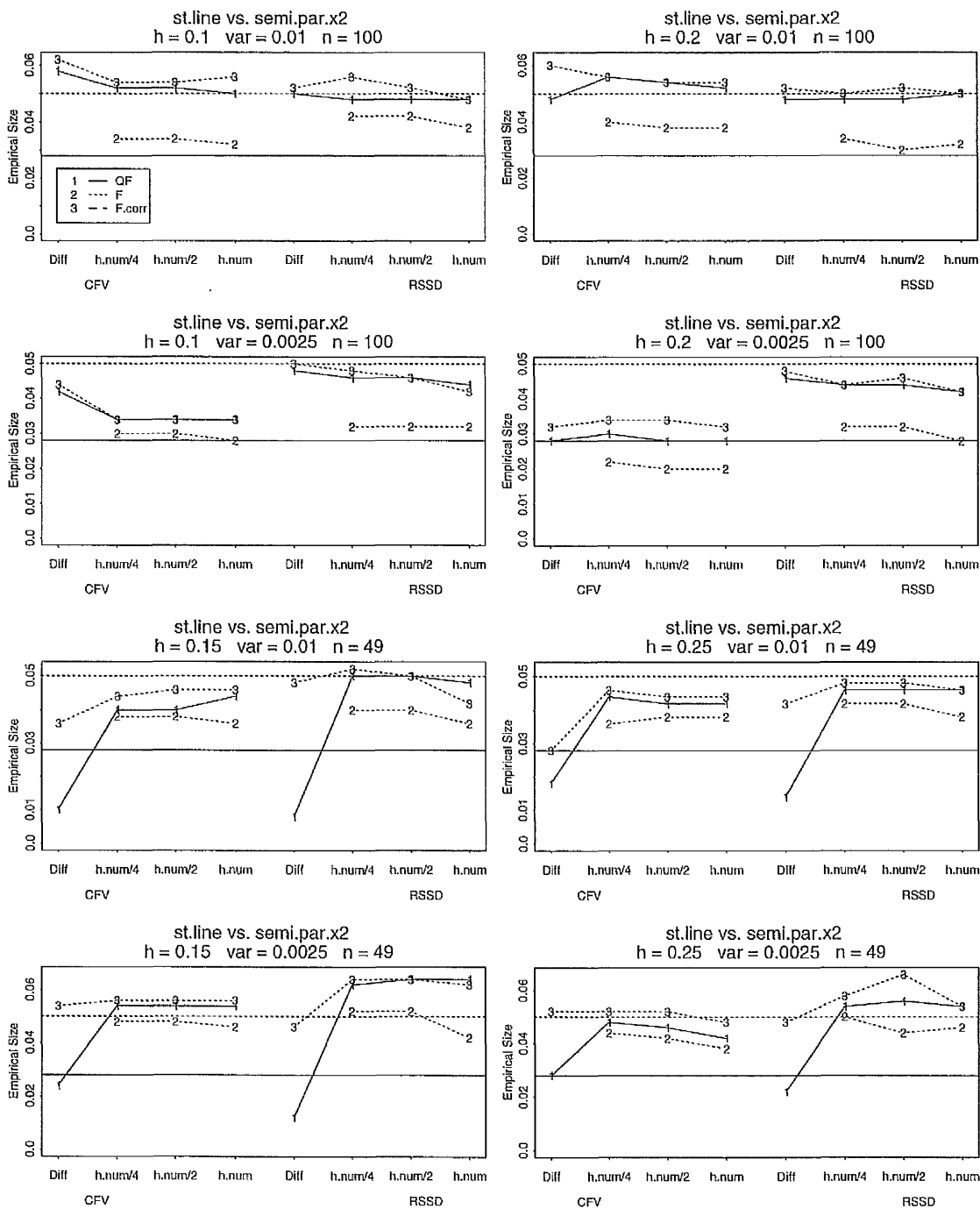
Data conditions	
design space	regular grid over $[0, 1]^2$
sample size	49 design points 100 design points
error distribution	Normal (zero mean)
error variance	0.0025 0.01
number of simulations	500
Test approaches	
model comparisons	2d.am vs. 2d.sm 1d.sm vs. 2d.am 1d.sm vs. semi.par.x1 semi.par.x1 vs. 2d.am st.line vs. semi.par.x2
smoothing parameter	0.15 & 0.25 when $n = 49$ 0.1 & 0.2 when $n = 100$
variance estimator ( $\sigma^2$ )	Diff. & RSS ( $\times 3$ )
comparisons of model fits	RSSD & CFV
reference distribution	QF, $F$ , corrected $F$

### Investigating size

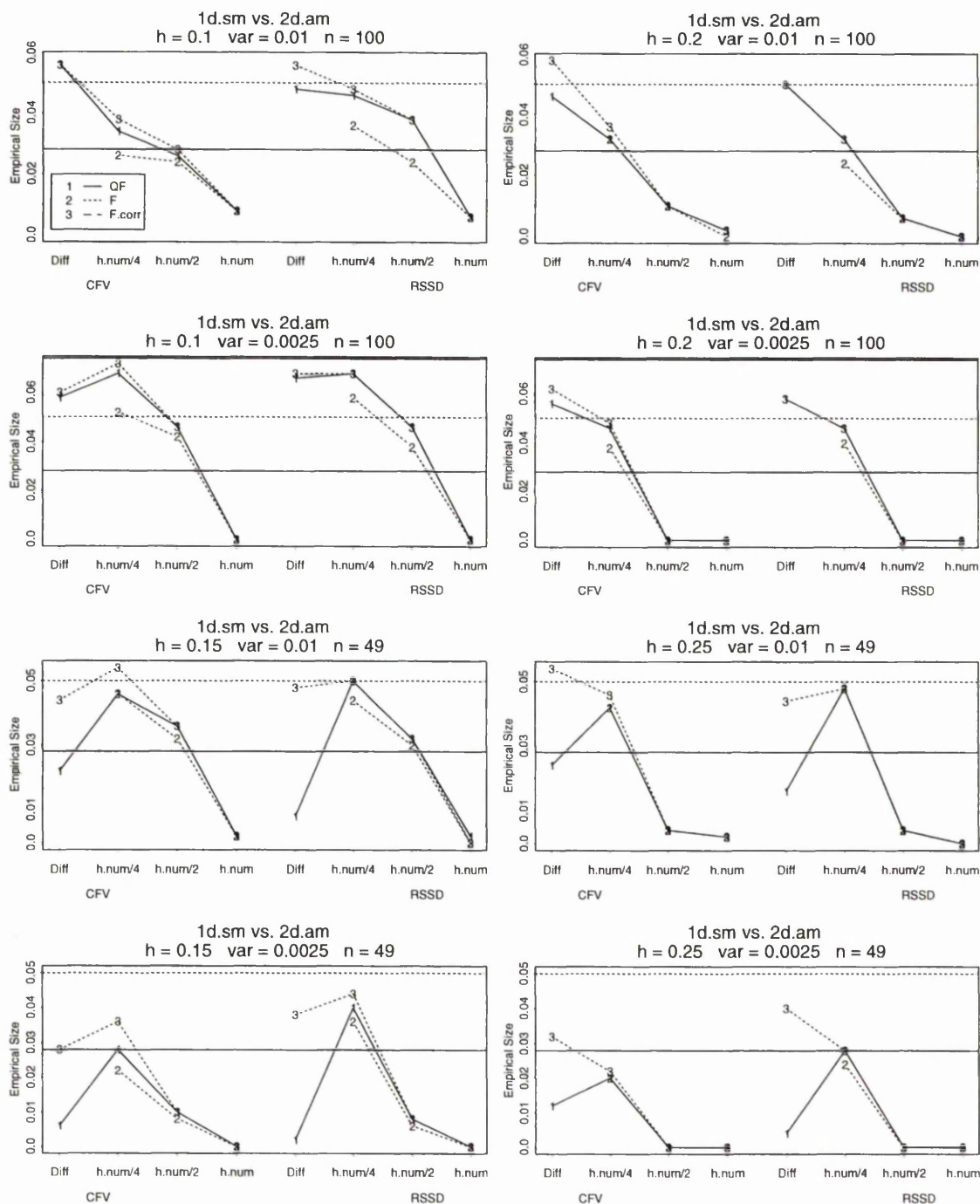
Empirical sizes corresponding to the settings in Table B.3 are listed in Tables B.4 - B.6 for each of the 24 combinations of test statistic (2), reference distribution (3) and estimate of  $\sigma^2$  (4). These results are summarised graphically in Figures 4.4-4.8, giving an indication of the performance of the size of each test under different settings. A 99% confidence from  $\text{Bin}(500, 0.05)$  is shown on each plot to aid in the assessment of the empirical sizes.

Figures 4.4-4.8 show that the combination of a *CFV* test statistics, scaled by the *difference based estimator* of  $\sigma^2$  and assessed using the two-moment corrected  $F$  distribution yield consistent results for all settings summarised in Table 4.2.

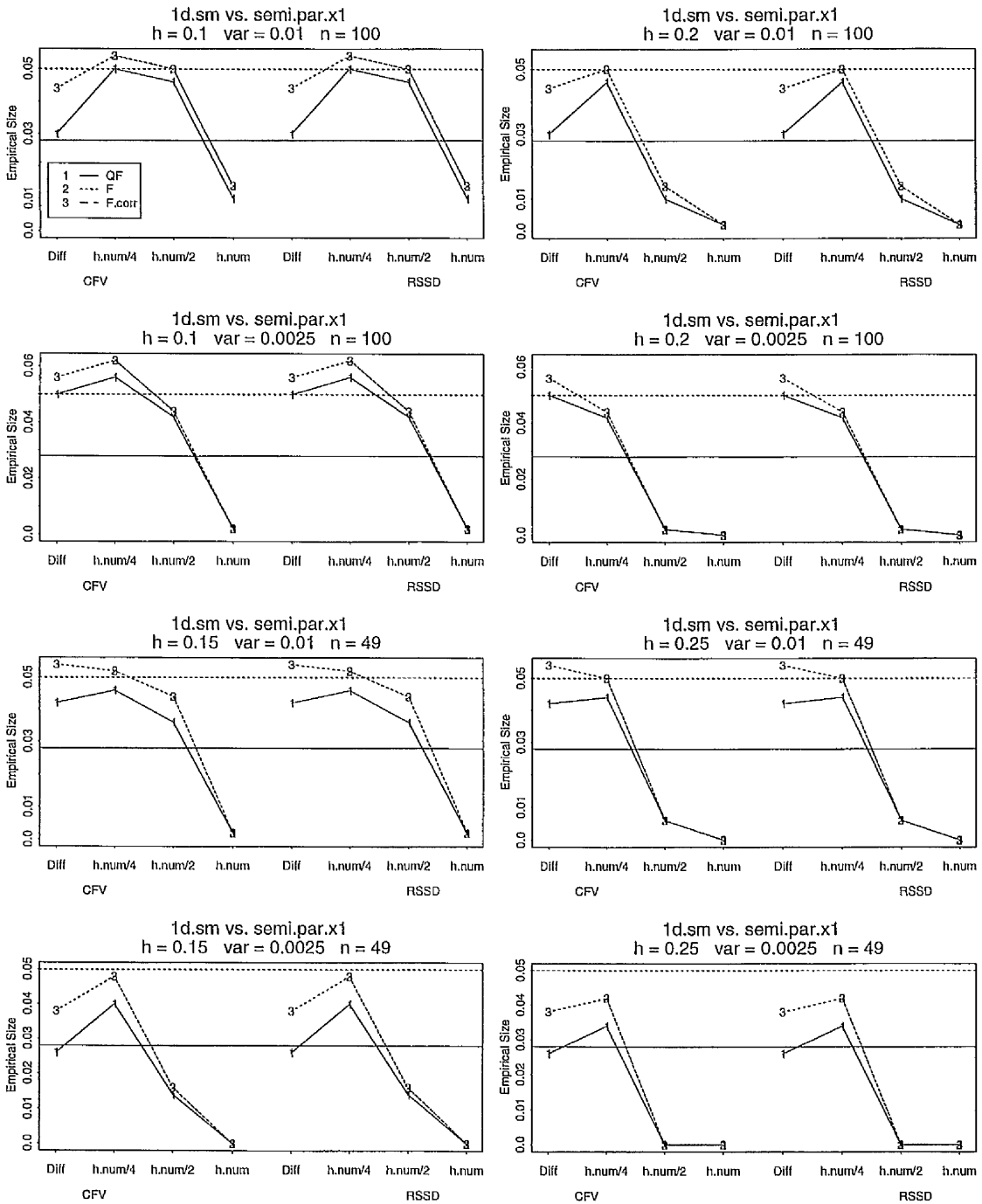
The results also highlight the importance of the choice of the smoothing parameter when the  $\hat{\sigma}_{RSS}^2$  is used to estimate  $\sigma^2$ . They show that tests using



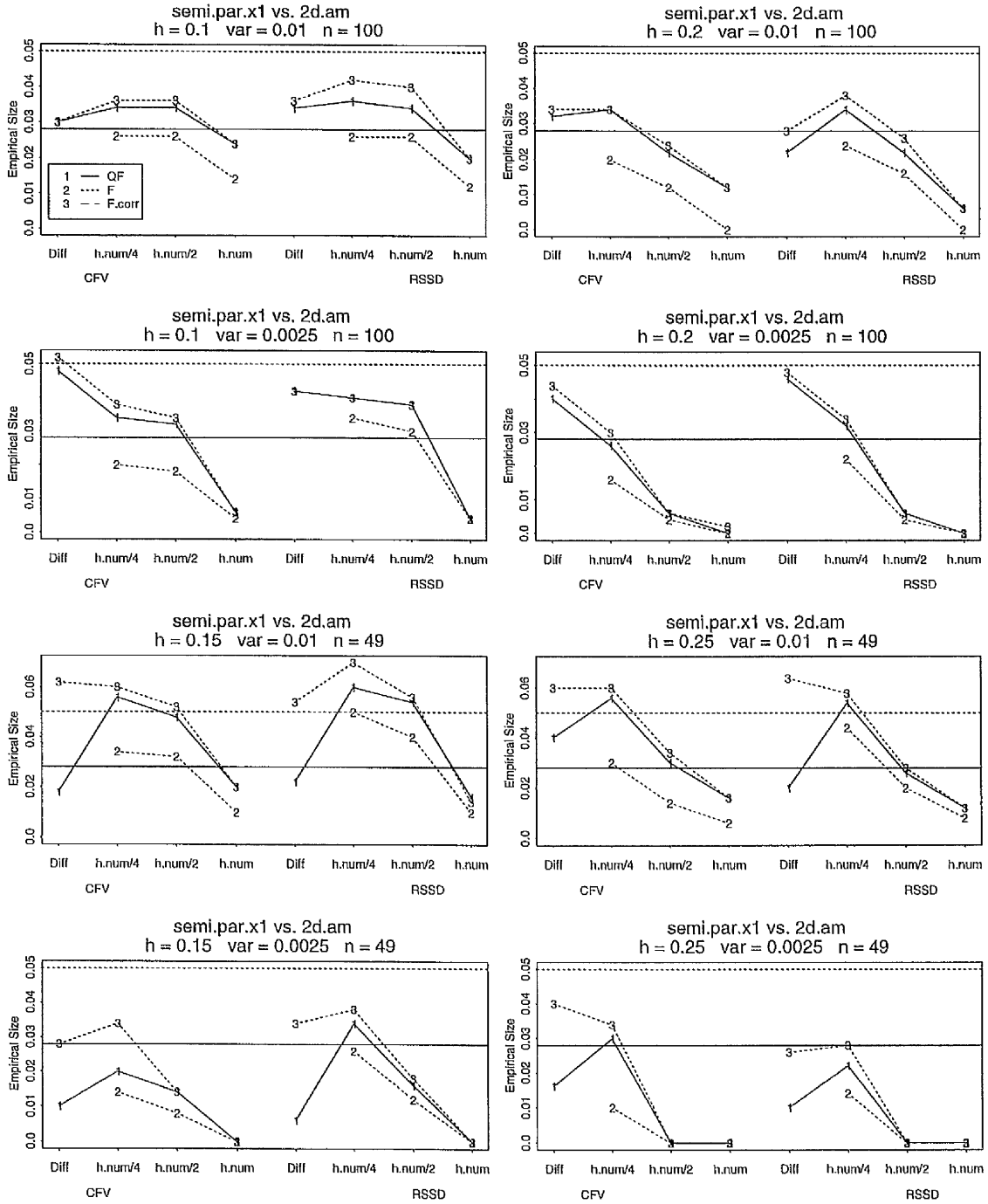
**Figure 4.4.** Size results of tests for model comparisons using 4 estimates of  $\sigma^2$ . 500 sets of observations from a *st.line* model were generated over a regular square grid. This (true) model fit was compared with a *semi.par* fit (linear in X) using 2 test statistics, and p-values calculated using 3 reference distributions, yielding the empirical sizes plotted.



**Figure 4.5.** Size results of tests for model comparisons using 4 estimates of  $\sigma^2$ . 500 sets of observations from a *1d.sm* model were generated over a regular square grid. This (true) model fit was compared with a *2d.am* fit using 2 test statistics, and p-values calculated using 3 reference distributions, yielding the empirical sizes plotted.

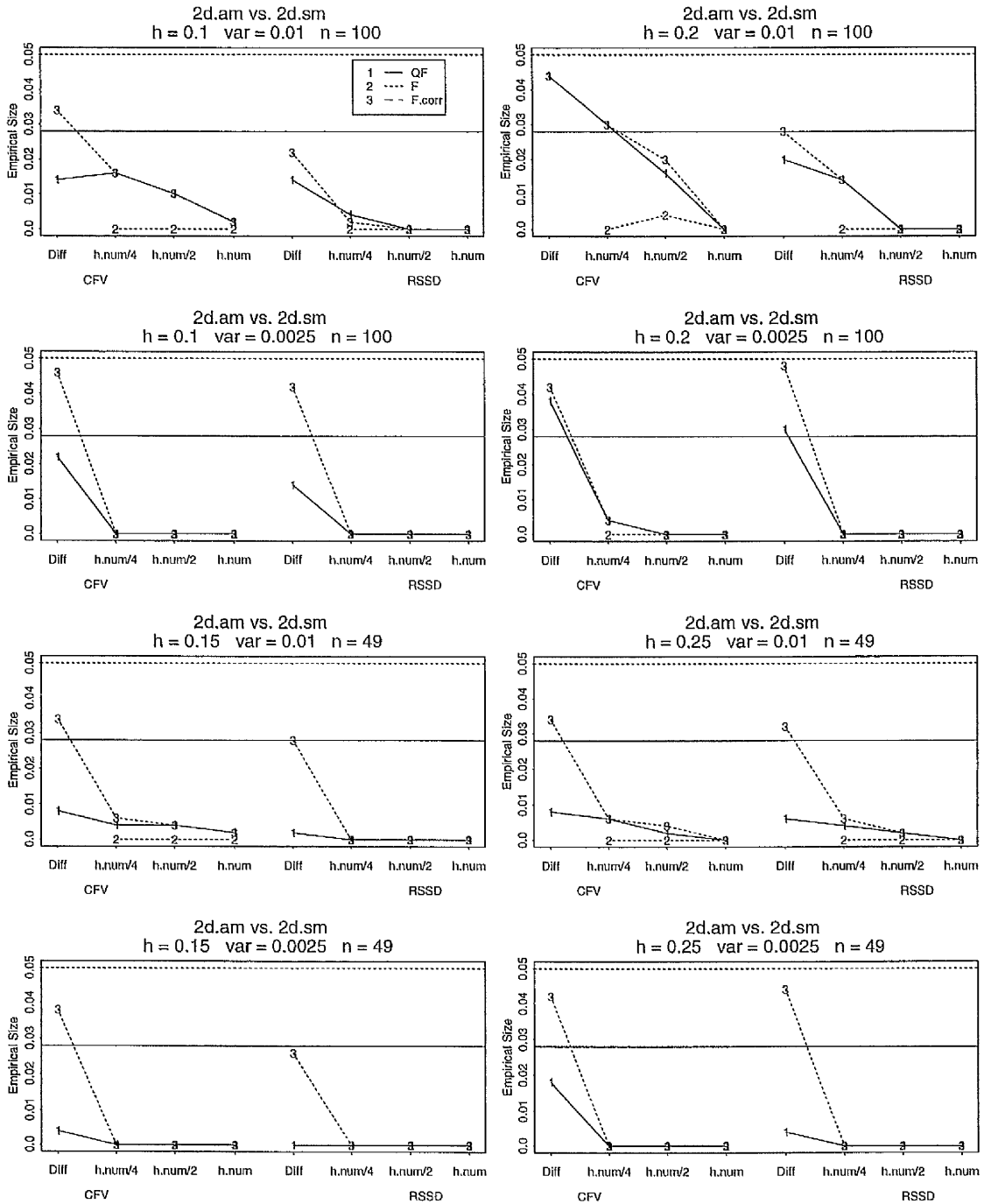


**Figure 4.6.** Size results of tests for model comparisons using 4 estimates of  $\sigma^2$ . 500 sets of observations from a *1d.sm* model were generated over a regular square grid. This (true) model fit was compared with a *semi.par* fit (linear in  $Z$ ) using 2 test statistics, and p-values calculated using 3 reference distributions, yielding the empirical sizes plotted.



**Figure 4.7.** Size results of tests for model comparisons using 4 estimates of  $\sigma^2$ . 500 sets of observations from a *semi.par* model were generated over a regular square grid. This (true) model fit was compared with a *2d.am* fit using 2 test statistics, and p-values calculated using 3 reference distributions, yielding the empirical sizes plotted.





**Figure 4.8.** Size results of tests for model comparisons using 4 estimates of  $\sigma^2$ . 500 sets of observations from a *2d.am* model were generated over a regular square grid. This (true) model fit was compared with a *2d.sm* fit using 2 test statistics, and p-values calculated using 3 reference distributions, yielding the empirical sizes plotted.

$\hat{\sigma}_{RSS}^2$  based on the same full model fit used in the model comparison statistic are conservative. That is, they are reluctant to suggest the full model provides a significant improvement to the fit of the reduced model. This is consistent with the earlier treatment of  $RSS$  based estimators of  $\sigma^2$  which showed that the effect of bias is to inflate the estimate which decreases the test statistic, thereby increasing its p-value and thus decreasing the empirical size observed over a number of simulations. This explains the observed effect of *undersmoothing* the full model (decreasing the smoothing parameter) for the purposes of estimating  $\sigma^2$  by  $\hat{\sigma}_{RSS}^2$  which is to increase the empirical sizes of the methods of inference so that they return values closer to the specified  $\alpha = 0.05$  significance level.

This property suggests that there is much sense in employing differenced based estimators of  $\sigma^2$  to perform model inference, at least in the case of a grid of design points (random designs will be considered in Section 4.3).

Interesting, too, is the role of the reference distribution. Despite its favourable performance when the known value of  $\sigma^2$  was used, the distribution derived from the distribution of quadratic forms does not capture the null distribution of the test statistic as well as the corrected F distribution. Similarly, and of greater consequence, is the poor performance of the standard F distribution. Given that this is the default distribution employed in practice and in statistical software such as S-Plus, these results highlight the care that should be taken in interpreting the results of standard analyses.

### Investigating power

Section 4.2.2 examined the performance of the size of the 24 combinations of two model fit comparisons, three reference distributions and four estimators of  $\sigma^2$ . This section completes the investigation for this case by considering the tests' power performances.

Using the same set of reduced and full models defined in Table 3.2, the following full models were found to deviate from the reduced model by a sufficient amount to provide informative results.

**semi.par.x1**  $m(X, Z) = 1 + X - 0.75 \exp(-0.5((X - 0.5)^2)/0.01) + \frac{1}{8}Z$

**semi.par.x2**  $m(X, Z) = \frac{1}{10}(1 + X - 0.75 \exp(-0.5((X - 0.5)^2)/0.01)) + Z$

$$\mathbf{2d.am} \quad m(X, Z) = 1 + X - 0.75 \exp(-0.5((X-0.5)^2)/0.01) + \frac{1}{6}(1 + Z + 0.75 \exp(-0.5((Z-0.5)^2)/0.01))$$

$$\mathbf{2d.sm} \quad \frac{1}{20}(-X + \Psi(X, Z, 0.3, 0.3, 0.1, 0.1, 0) + Z - \Psi(X, Z, 0.7, 0.7, 0.1, 0.1, 0))$$

where  $\Psi(\cdot)$  is, as before, the probability density function of a bivariate normal distribution.

Each empirical power listed in Table 4.3 is based on 500 sets of data simulated over regular grids. Only those approaches which were found to yield appropriate size results are included here. As such, only results based on two-moment corrected F distributions and  $\sigma^2$  estimators using differences or an undersmoothed ( $h/4$ ) RSS are included.

The results show that there is little to distinguish between the power results of the CFV and RSSD comparisons of model fits. Likewise there is only a marginal difference between the power results for difference based and undersmoothed RSS based estimators of  $\sigma^2$ . The former estimator does return higher powers for the *2d.am* vs. *2d.sm* comparison, a reflection of its ability to adapt to quite general (non-additive) regression surfaces. The two values of the smoothing parameter considered also returned similar empirical powers. Increasing the sample size had the expected effect of increasing the power of each test approach and setting considered.

### 4.3 Model Comparison Simulation Study: Random Design

The previous sections have examined the behaviour of the model comparison tests using data generated over regular grids of design points. We shall complete this study by considering the situation when design points *do not* comprise a grid of points, but rather when they are distributed randomly in the  $(X, Z)$  plane. Clearly, this type of design is closer to designs encountered in practice. However, differences in the designs lead to differences in the tests' performances. Therefore, before the simulation studies are described, properties of the model comparison tests over random designs are considered.

**Table 4.3.** Table listing power results from a simulation study over regular designs using estimated  $\sigma^2$ . Each result is with reference to a two moment corrected F distribution.

Red. Model	Full Model	n	h	Diff.		RSS - h.num/4	
				CFV	RSSD	CFV	RSSD
st.line	semi.par.x2	100	0.1	0.748	0.716	0.820	0.782
st.line	semi.par.x2	100	0.2	0.794	0.764	0.850	0.830
st.line	semi.par.x2	49	0.15	0.394	0.328	0.486	0.450
st.line	semi.par.x2	49	0.25	0.470	0.382	0.544	0.508
1d.sm	2d.am	100	0.1	0.944	0.910	0.964	0.948
1d.sm	2d.am	100	0.2	0.962	0.956	0.970	0.970
1d.sm	2d.am	49	0.15	0.612	0.562	0.694	0.676
1d.sm	2d.am	49	0.25	0.678	0.606	0.730	0.684
1d.sm	semi.par.x1	100	0.1	0.756	0.756	0.810	0.810
1d.sm	semi.par.x1	100	0.2	0.756	0.756	0.806	0.806
1d.sm	semi.par.x1	49	0.15	0.458	0.458	0.500	0.500
1d.sm	semi.par.x1	49	0.25	0.458	0.458	0.496	0.496
semi.par.x1	2d.am	100	0.1	0.492	0.492	0.546	0.564
semi.par.x1	2d.am	100	0.2	0.490	0.520	0.530	0.564
semi.par.x1	2d.am	49	0.15	0.196	0.212	0.232	0.270
semi.par.x1	2d.am	49	0.25	0.234	0.226	0.266	0.250
2d.am	2d.sm	100	0.1	0.728	0.694	0.716	0.726
2d.am	2d.sm	100	0.2	0.536	0.594	0.472	0.554
2d.am	2d.sm	49	0.15	0.140	0.110	0.062	0.030
2d.am	2d.sm	49	0.25	0.106	0.112	0.042	0.056

Section 3.2 highlighted the design adaptation property of local linear smooths, that is their asymptotic bias is independent of the distribution of the design points. Indeed, the performance of local linear regression over random designs is what recommends it over kernel (local constant) smoothers<sup>3</sup>. Although reassuring, the asymptotic behaviour can only act as a guide to the performance of model comparison tests. This is particularly so when models of different forms, each of which uses smoothers in different ways, are compared. Therefore, the *finite*

<sup>3</sup>Müller (1997) and Kneip & Engel (1996) have proposed modifications of the Nadaraya-Watson and Gasser-Müller kernel estimators which improve their properties under random designs.

*sample properties* of the model fit need to be investigated since these determine the performance of methods of inference.

It will be shown in this section that the symmetry of grid designs ensures the realisation of asymptotic properties in finite sample behaviour. This is why we have only mentioned the finite sample properties in passing up until this point. In the random design case, however, we will see that the finite and asymptotic bias properties differ. The results of the simulations over random designs need to be viewed with an understanding of these differences.

Another aspect which random designs introduce is the possibility of correlation amongst the covariates. It was noted in Section 3.2 that the asymptotic biases of semiparametric and additive model fits contain terms involving the joint distribution of the covariates. We therefore need to assess the effect of correlation amongst the covariates on the performance of the methods of inference.

Consider the simulated surface,  $m(X, Z) = 1 + X - 0.75 \exp(-0.5((X - 0.5)^2)/0.01)$ , first introduced in Section 2.7.2. Clearly this function produces a surface which is constant in the  $Z$  direction and a smooth curve in the  $X$  direction. This section addresses how the biases of three fitted models from the bivariate class (Section 3.2) applied to this simulated model, depend on the design points over which the model is simulated.

The expected fit of the models can be written as  $E(\hat{m}(\mathbf{x}, \mathbf{z})) = \mathbf{S}m(\mathbf{x}, \mathbf{z})$ , where  $\mathbf{S}$  is the smoothing matrix which yields the fitted model and  $(\mathbf{x}, \mathbf{z})$  are vectors containing the coordinates of the design points. Since the bias of the fitted model follows simply, it is therefore not necessary to add random errors to the regression function in order to investigate the finite sample bias properties of the different fits.

Three design spaces will be considered: a regular grid, an irregular grid and a random design as illustrated in Figure 4.9. The first panel shows a regular  $n \times n$  grid consisting of  $n^2 = 100$  equally spaced points over the plane. The second panel represents an irregular grid which is produced by crossing two sets of 10 ( $n$ ) random variables generated from a uniform distribution. The random design is produced by pairing two independent sets of  $n = 100$  uniformly distributed random variables, one realisation of which is shown in the third panel of Figure 4.9.

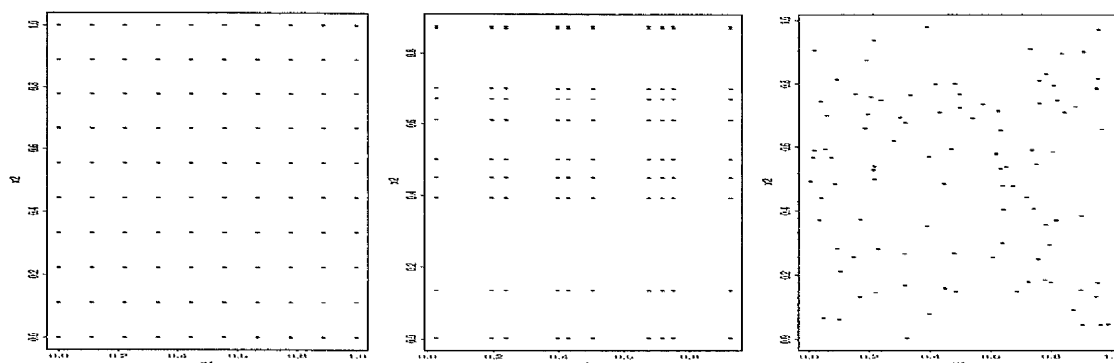


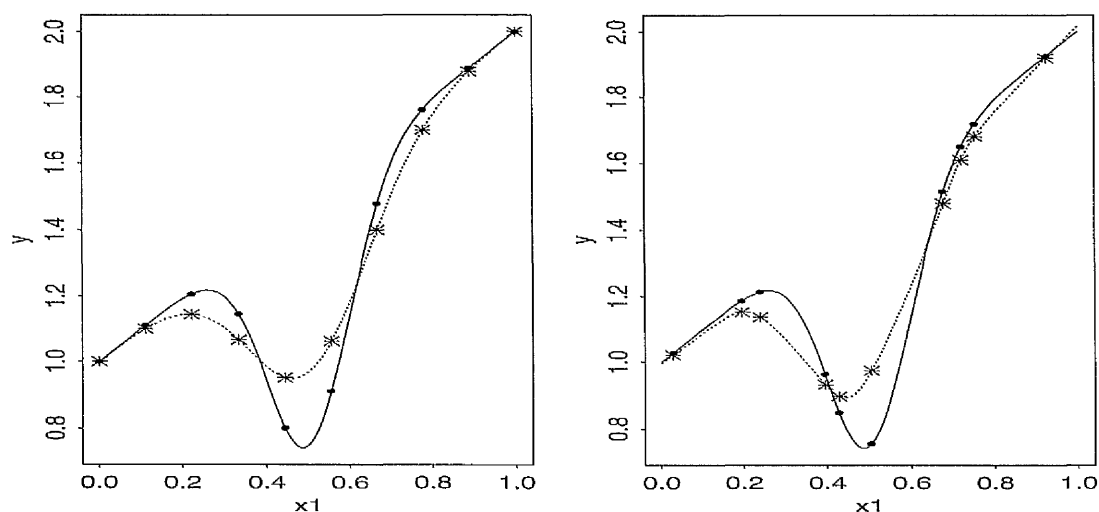
Figure 4.9. Three covariate design spaces.

Using the three designs displayed in Figure 4.9, data were generated from the regression function  $m(\cdot, \cdot)$  above. For each of the three designs, the expected values of a univariate local linear smooth, a bivariate local linear smooth and a two dimensional additive model were calculated.

The left and right panels of Figure 4.10 display, for the regular and irregular grid respectively, the true curve (solid line), the true values at the design points ( $\bullet$ ), the expected value of the univariate fit (dotted line), the expected value of the bivariate fit ( $\times$ ) at the design points and the expected value of the two dimensional additive fit at the design points ( $+$ ). Figure 4.11 shows the results using a random design. The true curve and the expected values of the univariate, additive and bivariate fit are shown as in Figure 4.10. In Figure 4.11, however, the expected values of the bivariate fit ( $\times$ ) and the expected values of the two dimensional additive fit ( $+$ ) are shown in separate panels. Because there is no effect in the  $Z$  direction, the two dimensional surfaces are projected onto the  $(X, Y)$  plane.

Figure 4.10 shows that the expected values (and therefore the finite sample bias) of the three fitted models are equivalent under both the regular grid and the irregular grid, whereas Figure 4.11 shows that this is not the case for the random design. The discrepancy between the biases of the fits over the random design is of particular interest as it shows that the bias of the two fits will not ‘cancel’ in the CFV statistic when the design space is a random design.

These numerical results for the grid designs can be expressed as:

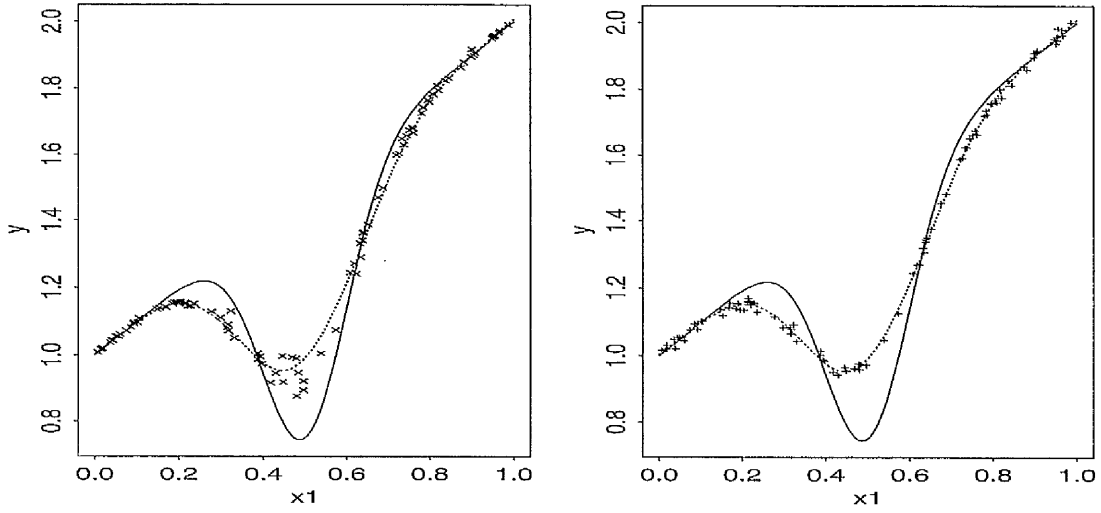


**Figure 4.10.** Expectations of fitted values under different nonparametric models for the *regular grid* (first panel) and the *irregular grid* (second panel). In each panel the true model is displayed as a solid line and the expectation of the *univariate smooth* is displayed as a dotted line. The expected values of the *bivariate smooth* are shown as  $\times$  and the expected values of a *two dimensional additive model* are shown as  $+$ .

$$S_U^G m(\mathbf{x}) = S_A^G m(\mathbf{x}) = S_B^G m(\mathbf{x}), \quad (4.1)$$

where  $S_U^G$ ,  $S_A^G$  and  $S_B^G$  are the  $n^2 \times n^2$  smoothing matrices evaluated over the grid design and which generate the univariate, the additive and the bivariate fits respectively. Expression 4.1 shows that the smoothing matrices inherit properties from the grid design which yield the equivalent finite sample biases observed here. In particular, it is the constant distribution of points in the  $Z$  direction which ensures that the weights of the smoothing matrices behave in this way, i.e. combine to return an expected surface with a no-effect in the  $Z$  direction. Note that these properties can also be shown algebraically using the definitions of the smoothing weights given in Section 3.2.

Conversely, the results of Figure 4.11 show that smoothing matrices calculated over random designs do not share these properties. In other words, the lack of constancy in the distribution of design points in the plane results in smoothing



**Figure 4.11.** Expectations of fitted values under different nonparametric models for the *random* design. In each panel the true model is displayed as a solid line and the expectation of the *univariate smooth* is displayed as a dotted line. The expected values of the *bivariate smooth* are shown in the first panel as  $\times$  and the expected values of a *two dimensional additive model* are shown in the second panel as  $+$ .

weights for the additive and bivariate smooth fits which return non-constant expected surfaces in the  $Z$  direction, even though there is no effect in this direction in the regression surface. The result is that the finite sample biases of these fits will not cancel with each other or with that from the univariate smooth.

These numerical results can be understood more fully by using several properties of the local linear regression smoother together with Taylor's theorem to derive expressions for the exact finite sample biases of the fits. First consider a univariate smoother with a smoothing matrix whose  $(i, j)$ th element is the weight  $w_j(X_i)$ . Assuming that the regression function  $m(\cdot)$  which underlies the data is a univariate function with a continuous second derivative, Taylor's theorem states:

$$E(\hat{m}(X_i)|\mathbf{x}) = \sum_{j=1}^n w_j(X_i) E(Y_j|X = X_j)$$



$$\begin{aligned}
&= \sum_{j=1}^n w_j(X_i) m(X_j) \\
&= m(X_i) \sum_{j=1}^n w_j(X_i) + m'(X_i) \sum_{j=1}^n (X_j - X_i) w_j(X_i) + \\
&\quad \frac{1}{2} \sum_{j=1}^n (X_j - X_i)^2 w_j(X_i) m''(\theta_j)
\end{aligned}$$

where  $(\theta_j - X_i)(\theta_j - X_j) \leq 0$ . Noting that local linear regression has the property that it preserves *linear functions*, i.e.  $\sum_{j=1}^n w_j(X_i) = 1$  and  $\sum_{j=1}^n w_j(X_i) X_j = X_i$ , an exact expression for the bias is therefore

$$E(\hat{m}(X_i)|\mathbf{x}) - m(X_i) = \frac{1}{2} \sum_{j=1}^n (X_j - X_i)^2 w_j(X_i) m''(\theta_j). \quad (4.2)$$

Similar results can be derived for an additive model fit with smoothing matrix weights given by  $w_{Aj}(X_i, Z_i)$ . That is, when the underlying regression function is a bivariate additive surface the finite sample bias is given by:

$$\begin{aligned}
E(\hat{m}(X_i, Z_i)|\mathbf{x}, \mathbf{z}) - m(X_i, Z_i) &= \frac{1}{2} \sum_{j=1}^n (X_j - X_i)^2 w_{Aj}(X_i, Z_i) m''_1(\theta_{1j}) + \\
&\quad \frac{1}{2} \sum_{j=1}^n (Z_j - Z_i)^2 w_{Aj}(X_i, Z_i) m''_2(\theta_{2j}).
\end{aligned}$$

where  $(\theta_{1j} - X_i)(\theta_{1j} - X_j) \leq 0$  and  $(\theta_{2j} - Z_i)(\theta_{2j} - Z_j) \leq 0$ .

Of interest to the present discussion is how these finite sample biases compare when applied to the same data. If the underlying regression function is univariate then  $m''_2(\cdot)$  is clearly zero and the bias of the additive fit reduces to  $\frac{1}{2} \sum_{j=1}^n (X_j - X_i)^2 w_{Aj}(X_i, Z_i) m''_1(\theta_{1j})$ . The form is very similar to Equation 4.2 with the notation suggesting that differences lie in the weights of the smoothing matrices and the unknown  $\theta$ 's.

Another form of Taylor's theorem, however, states that  $\theta_j$  in Equation 4.2 satisfies  $\frac{1}{2} m''(\theta_j)(X_j - X_i)^2 = \int_{X_i}^{X_j} m''(t)(X_j - t) dt$ . Incorporating this definition

into Equation 4.2 yields

$$E(\hat{m}(X_i)|\mathbf{x}) - m(X_i) = \sum_{j=1}^n w_j(X_i) \int_{X_i}^{X_j} m''(t)(X_j - t)dt$$

and, similarly, the additive fit's bias when the regression function is univariate reduces to:

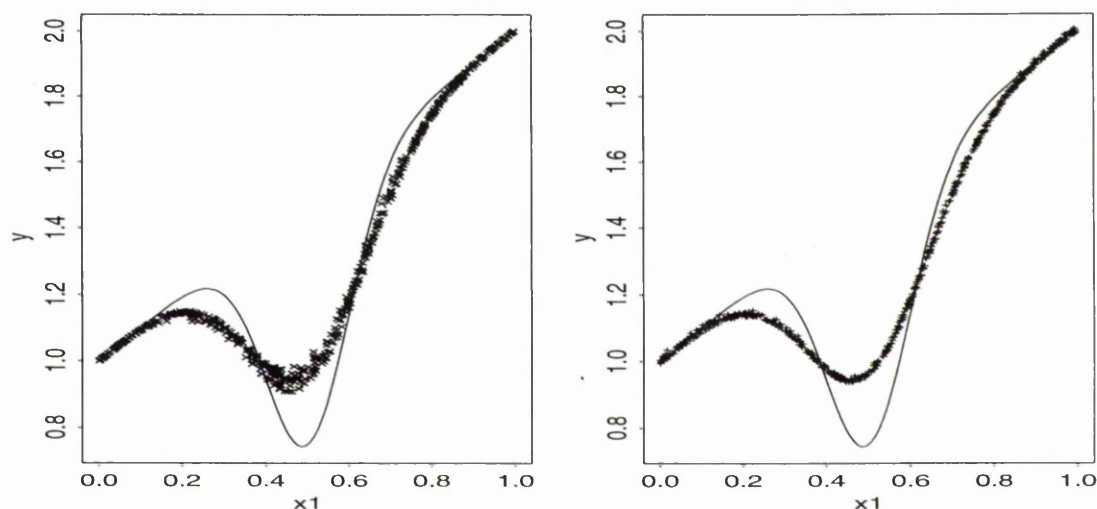
$$E(\hat{m}(X_i, Z_i)|\mathbf{x}, \mathbf{z}) - m(X_i, Z_i) = \sum_{j=1}^n w_{Aj}(X_i, Z_i) \int_{X_i}^{X_j} m''(t)(X_j - t)dt.$$

Hence, the finite sample biases of these two fits to the same data differ only with respect to the weights in their respective smoothing matrices.

Corresponding expressions can be derived for the semiparametric and bivariate smooth fits. Comparisons amongst these finite sample bias expressions reveal that they are also equivalent except for the weights of the smoothing matrices when the regression function is of the same form as the 'reduced' model fit. The properties of smoothing matrices over regular grids of design points noted above lead to identical finite sample biases, as seen in Figure 4.10.

Given, therefore, that the differences in model biases over random designs seen in Figure 4.11 are a symptom of the finite sample realisations, these discrepancies in expected fits should decrease as the sample size increases. Figure 4.12 was produced under identical conditions to Figure 4.11 yet with 625 simulated design points as opposed to 100. It shows that the discrepancies between the biases of the univariate fit and both the additive and the bivariate smooth are less for this larger sample size. Of course, asymptotically the expectation of the fitted additive and bivariate surface will be constant in the  $Z$  direction and equivalent to the univariate expectation, however these finite sample properties are clearly relevant to methods of inference.

Given that the biases of the two models no longer cancel when design points form a random configuration, we need to consider the impact on the test procedures described previously. Recall that model comparisons were based on either direct comparisons of the fitted values (CFV) or differences in the RSS of the two



**Figure 4.12.** Expectations of fitted values under different nonparametric models for the *random* design using 625 simulated points. In each panel the true model is displayed as a solid line and the expectation of the *univariate smooth* is displayed as a dotted line. The expected values of the *bivariate smooth* are shown in the first panel as  $\times$  and the expected values of a *two dimensional additive model* are shown in the second panel as  $+$ .

models. Given that the two model fits to be compared are given by the smoothing matrices  $\mathbf{S}_R$  and  $\mathbf{S}_F$ , representing reduced and full models respectively, and that the observed data are of the form  $\mathbf{y} = m(\mathbf{x}, \mathbf{z}) + \boldsymbol{\varepsilon}$ , then the CFV model comparison statistic was shown in Section 3.4 to be:

$$\begin{aligned}
 CFV &= \mathbf{y}^T (\mathbf{S}_R - \mathbf{S}_F)^T (\mathbf{S}_R - \mathbf{S}_F) \mathbf{y} \\
 &= m(\mathbf{x}, \mathbf{z})^T (\mathbf{S}_R - \mathbf{S}_F)^T (\mathbf{S}_R - \mathbf{S}_F) m(\mathbf{x}, \mathbf{z}) + \boldsymbol{\varepsilon}^T (\mathbf{S}_R - \mathbf{S}_F)^T (\mathbf{S}_R - \mathbf{S}_F) \boldsymbol{\varepsilon} \\
 &= (\mathbf{b}_R - \mathbf{b}_F)^T (\mathbf{b}_R - \mathbf{b}_F) + \boldsymbol{\varepsilon}^T (\mathbf{S}_R - \mathbf{S}_F)^T (\mathbf{S}_R - \mathbf{S}_F) \boldsymbol{\varepsilon}
 \end{aligned} \tag{4.3}$$

This expression shows that the effect of the discrepancies in bias between the reduced and the full model is to inflate the value of CFV. Therefore, test procedures which assume that the biases cancel will return an artificially low p-value which will lead to the rejection of the null hypothesis (when it is true)

more frequently than the specified significance level. Therefore, we expect our simulation results under the null (reduced) model to return empirical p-values significantly greater than the conventional 0.05 level used.

A similar expression can be derived for the RSSD based test:

$$\begin{aligned}
 RSSD &= \mathbf{y}^T[(\mathbf{I} - \mathbf{S}_R)^T(\mathbf{I} - \mathbf{S}_R) - (\mathbf{I} - \mathbf{S}_F)^T(\mathbf{I} - \mathbf{S}_F)]\mathbf{y} \\
 &= \mathbf{m}(\mathbf{x})^T[(\mathbf{I} - \mathbf{S}_R)^T(\mathbf{I} - \mathbf{S}_R) - (\mathbf{I} - \mathbf{S}_F)^T(\mathbf{I} - \mathbf{S}_F)]\mathbf{m}(\mathbf{x}) + \\
 &\quad \varepsilon^T[(\mathbf{I} - \mathbf{S}_R)^T(\mathbf{I} - \mathbf{S}_R) - (\mathbf{I} - \mathbf{S}_F)^T(\mathbf{I} - \mathbf{S}_F)]\varepsilon - \\
 &\quad 2\varepsilon^T[(\mathbf{I} - \mathbf{S}_R)^T(\mathbf{I} - \mathbf{S}_R) - (\mathbf{I} - \mathbf{S}_F)^T(\mathbf{I} - \mathbf{S}_F)]\mathbf{m}(\mathbf{x}) \\
 &= \mathbf{b}_R^T\mathbf{b}_R - \mathbf{b}_F^T\mathbf{b}_F + \varepsilon^T[(\mathbf{I} - \mathbf{S}_R)^T(\mathbf{I} - \mathbf{S}_R) - (\mathbf{I} - \mathbf{S}_F)^T(\mathbf{I} - \mathbf{S}_F)]\varepsilon - \\
 &\quad 2\varepsilon^T[(\mathbf{I} - \mathbf{S}_R)^T\mathbf{b}_R - (\mathbf{I} - \mathbf{S}_F)^T\mathbf{b}_F].
 \end{aligned} \tag{4.4}$$

Since the full model allows more flexibility in the fit, the expected values of the reduced model will lie further away from the true model than the full model's expected fit. Hence  $\mathbf{b}_R^T\mathbf{b}_R > \mathbf{b}_F^T\mathbf{b}_F$ . Of course, the cross product term remains, but its expectation is zero and therefore on average its influence will cancel out.

In the case of the two dimensional semiparametric and additive model, we saw in Section 3.2 how the joint distribution of the covariates appears in the expression for asymptotic biases of the fits when the two covariates were not independent. This property suggests that even for large samples, discrepancies between the biases may affect the performance of model comparisons involving additive and semiparametric fits.

To explore the impact of dependency between the covariates on the bias properties, 500 sets of covariates were simulated from the bivariate normal distributions ( $\mu_1 = \mu_2 = 0.5$ ,  $\sigma_1 = \sigma_2 = 0.15$ ,  $n = 100$ ) with correlation coefficients  $\rho = 0, 0.2, 0.4, 0.6, 0.8$ . For each of these simulated designs smoothing matrices were calculated for four model fits at a range of smoothing parameters. Amongst these models the following comparisons were made:

- ◊ additive vs. bivariate
- ◊ semiparametric vs. additive

- ◇ univariate vs. additive
- ◇ univariate vs. semiparametric

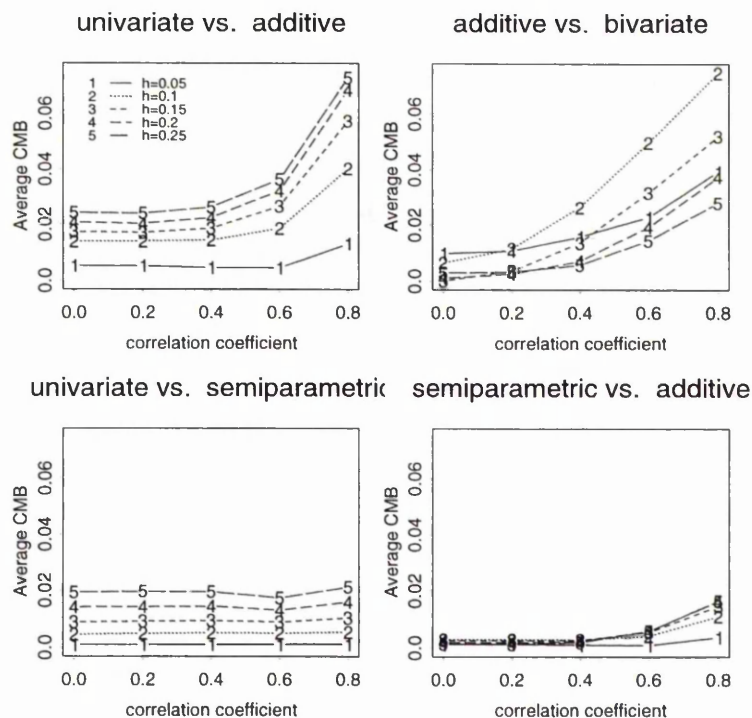
In each case a *comparison of model biases*,  $CMB = (\mathbf{b}_R - \mathbf{b}_F)^T(\mathbf{b}_R - \mathbf{b}_F)$ , which appears in the CFV expression above, is calculated using the true regression function corresponding to ‘reduced’ model. Figure 4.13 displays the average  $CMB$  (over the 500 simulated designs) for each each combination of  $\rho$  and  $h$ .

These results demonstrate how the joint distribution of the design points impacts the CFV statistic through the bias term. The effect is most noticeable when the additive model is compared with a univariate and a bivariate smooth fit. This is consistent with the asymptotic bias results. There is also some evidence that the correlation effect which appears in both the additive and the semiparametric model biases cancels when these fits are compared. Although asymptotically there is an extra term in the semiparametric bias, even when the regression function is univariate, these results indicate that this is not related to the correlation coefficient of the bivariate normal.

This section has illustrated and explored some of the bias properties of the nonparametric model fits when the designs are random configurations. These properties have obvious consequences for model comparisons although there are other factors at work, such as the reference distributions employed and the estimation of  $\sigma^2$ . Sections 4.3.1 and 4.3.2 explore methods of inference for independent random designs while Section 4.3.3 investigates the effect of correlation on the best test procedure for random designs identified by the simulation study of Section 4.3.2.

### 4.3.1 Random design: $\sigma^2$ known

We start our investigation of the performance of the tests comparing model fits over random designs by focusing on the CFV and RSSD statistics through using the known value of  $\sigma^2$  in the test procedures. The tests and conditions considered in this section are summarised in Table 4.4 and listed in full in Table B.7. Note the introduction of three random configurations of design points. That is, in addition to the random errors about the true regression function, each simulated



**Figure 4.13.** Comparisons of Model Biases (CMB) averaged over 500 simulated designs ( $n = 100$ ) from bivariate normal distributions with varying levels of correlation.

data set differs with respect to the design points over which it is calculated. To aid in the evaluation of factors affecting the performance of the test methods, the same set of simulated designs was used for each combination of *models compared* and *sample size*. That is, each block of four rows in Table B.7 uses the same set of 500 simulated designs. Indeed each pair of rows in Table B.7 used the same simulated data, since the settings differ only in the value of the smoothing parameter used.

Tables B.8 - B.10 list the empirical sizes of each setting for the three types of random design respectively (under the conditions listed in Table B.7). Figures 4.14-4.16 present these results graphically, the results in each plot being based on the same set of 500 simulated designs.

A comparison of each of Figures 4.14-4.16 with Figure 4.2 reveals the dramatic effect of using random designs rather than fixed regular grids. There are now

**Table 4.4.** Summary of the settings and the test procedures used in the simulations of Section 4.3.1.

Data conditions	
design space	uniform-random over $[0, 1]^2$ bivariate normal ( $\rho = 0$ ) $\mu_1 = \mu_2 = 0.5$ , $\sigma_1 = \sigma_2 = 0.15$ bivariate normal ( $\rho = 0.5$ ) $\mu_1 = \mu_2 = 0.5$ , $\sigma_1 = \sigma_2 = 0.15$
sample size	49 design points 100 design points
error distribution	Normal (zero mean)
error variance	0.01 0.04
number of simulations	500
Test approaches	
model comparisons	2d.am vs. 2d.sm 1d.sm vs. 2d.am 1d.sm vs. semi.par.x1 semi.par.x1 vs. 2d.am st.line vs. semi.par.x2
smoothing parameter	0.05, 0.15
variance estimator ( $\sigma^2$ )	NONE (true value used)
comparisons of model fits	RSSD & CFV
reference distribution	QF, $\chi$ , corrected $\chi^2$

far more instances of the tests returning sizes outwith values consistent with a  $\alpha = 0.05$  significance level. Care must therefore be taken to determine the correct approach to model inference when a random design is present.

A striking feature of these results is the difference between the model comparisons. It is therefore helpful to summarise each of these cases separately.

**st.line vs. semi.par.x2:** in this ‘bias-free’ case it is not surprising to see most tests returning acceptable values. Problems are noted when RSSD comparisons are employed with a smaller number of design points distributed uniformly, especially at the smaller smoothing parameter.

**1d.sm vs. 2d.am:** this case poses severe challenges to the methods of inference.

Across the three design configurations, a CFV comparison with reference

to an unadjusted  $\chi^2$  distribution returns sensible results, but only at the higher level of error variance. All the other tests and approaches reject the null hypothesis far too readily.

**1d.sm vs. semi.par.x1:** here problems arise with all test procedures at the lower error variance using the higher smoothing parameter. At other settings all of the test procedures return consistent results.

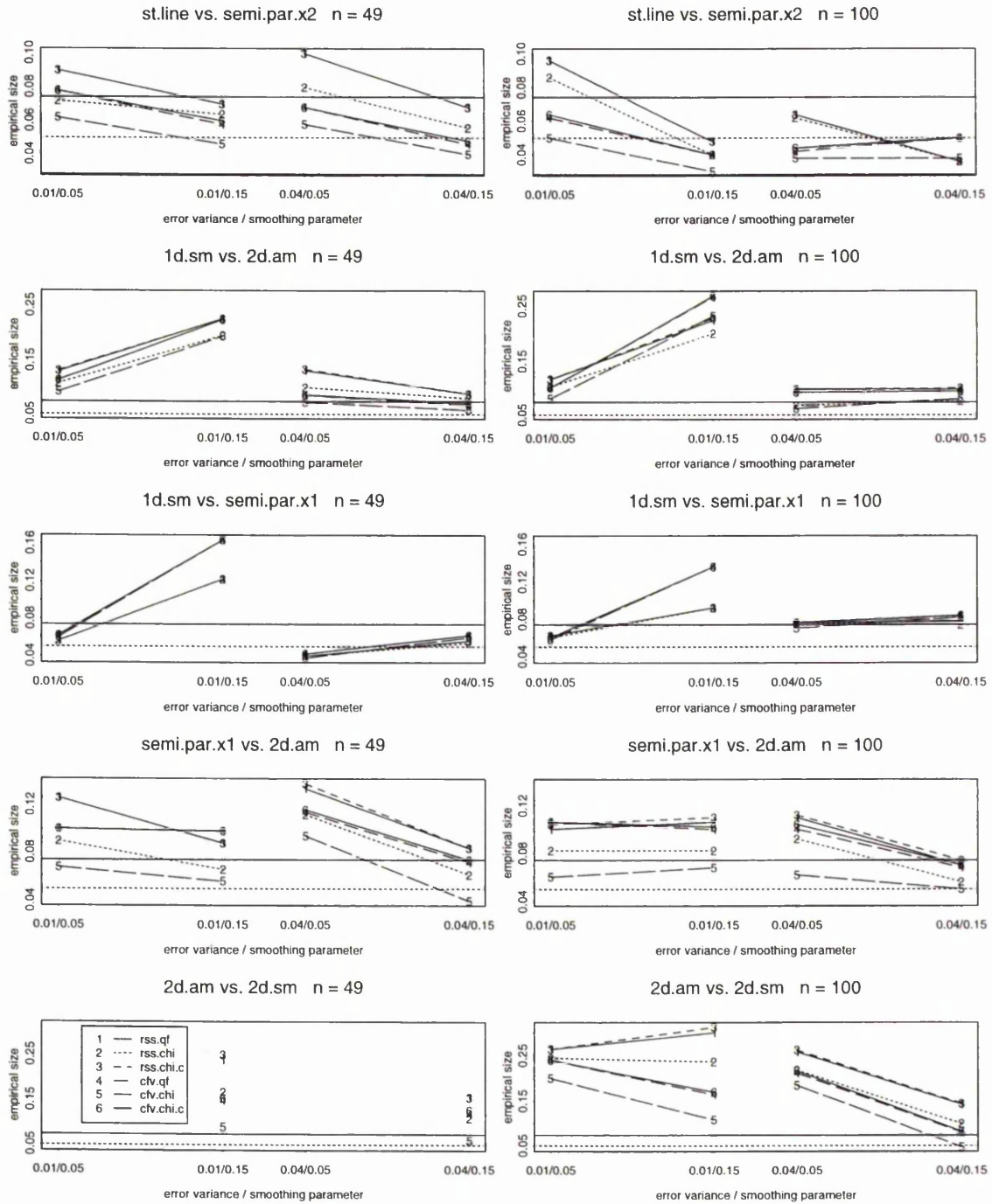
**semi.par.x1 vs. 2d.am:** there is clear evidence in this comparison to favour a CFV comparison with reference to a  $\chi^2$  distribution over other approaches. This is the only approach which yields consistent results in nearly all settings.

**2d.am vs. 2d.sm:** this comparison also presents major challenges. The only approach which returns acceptable sizes is, once again, a CFV comparison with reference to a  $\chi^2$  distribution, but only at the higher smoothing parameter and the higher error variance (except in the uncorrelated bivariate design where it is consistent at both values of error variance).

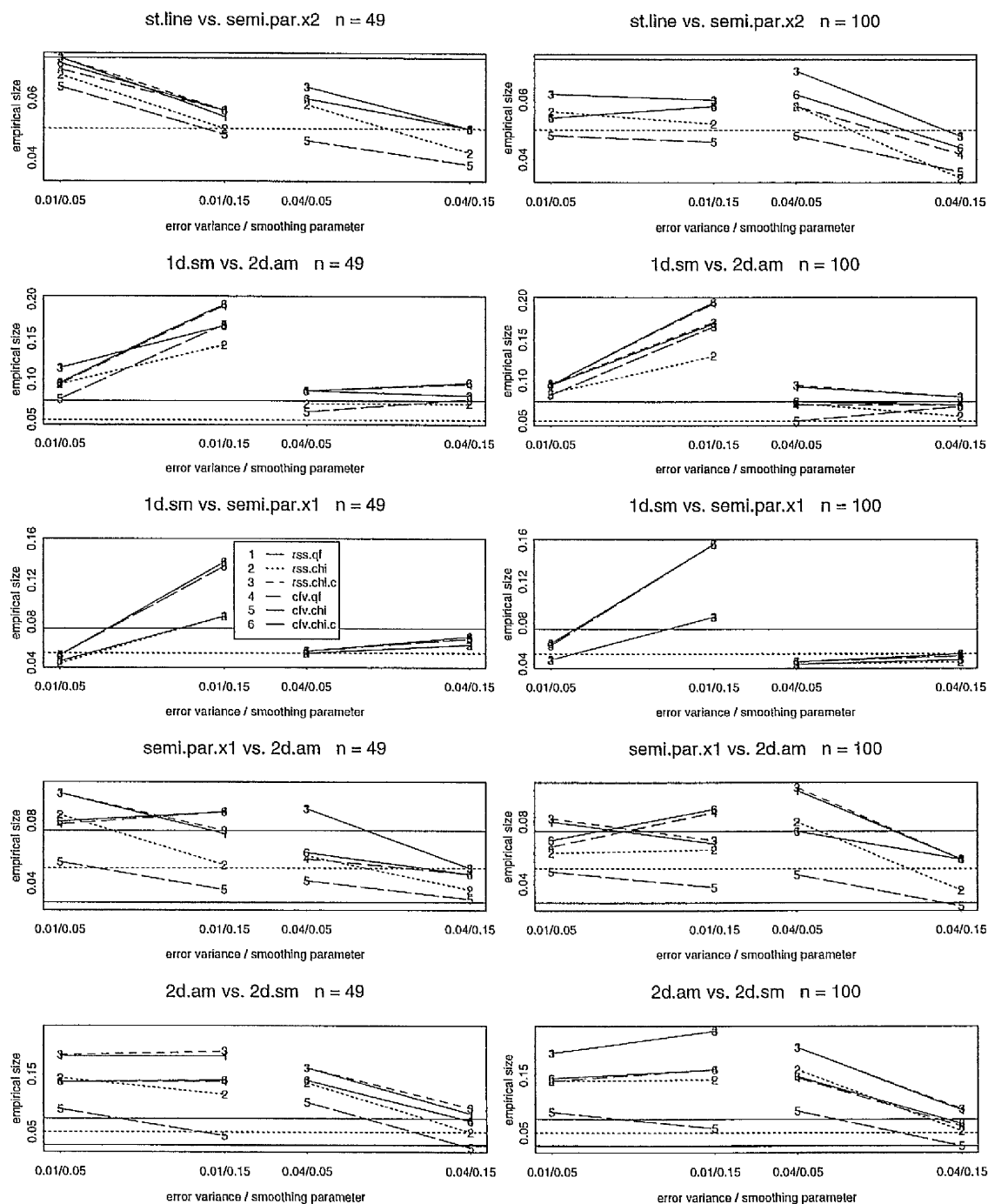
One property worth noting here is a comparison of the additive vs. bivariate smooth results for the two bivariate normal designs: one with zero correlation, the other with correlation coefficient of 0.5. We noted earlier the effects of correlated covariates on the bias properties of additive model fits. Here the introduction of correlation into the design dramatically increased the empirical sizes of each of the tests, especially at the smaller value of  $\sigma^2$ . This property will be investigated further in Section 4.3.3 where estimates of  $\sigma^2$  are used and several correlation coefficients are used to generate the random designs.

This section has highlighted clearly the complications caused by the finite sample bias properties of smoothers generated over random designs. By using the known value of  $\sigma^2$  these simulations have isolated the effect on the comparison of model fits statistics CFV and RSSD. However, the estimators of  $\sigma^2$ , which are in practice necessary, will also be influenced by the finite sample bias properties and therefore the next section will incorporate the estimation of  $\sigma^2$  into the test procedures.

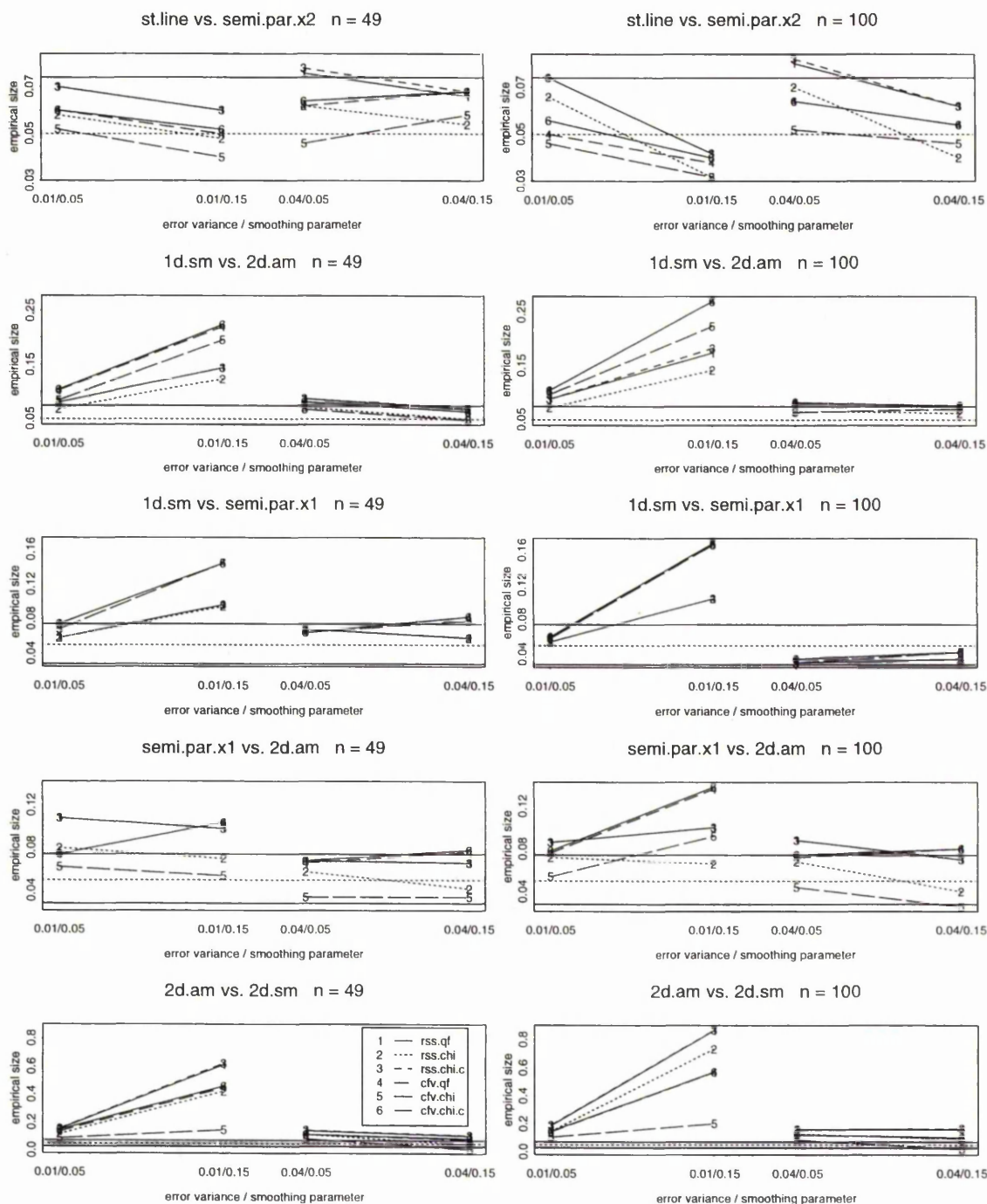




**Figure 4.14.** Empirical sizes of tests of model comparisons using known  $\sigma^2$ . Design points are realisations of *uniform random* designs. The results listed are the proportion of 500 simulated p-values less than 0.05 under the reduced model. The solid horizontal lines (when shown) indicate a 99% confidence interval from  $\text{Bin}(500, 0.05)$ .



**Figure 4.15.** Empirical sizes of tests of model comparisons using known  $\sigma^2$ . Design points are realisations of *bivariate normal* designs with  $(\rho = 0)$ . The results listed are the proportion of 500 simulated p-values less than 0.05 under the reduced model. The solid horizontal lines (when shown) indicate a 99% confidence interval from  $\text{Bin}(500, 0.05)$ .



**Figure 4.16.** Empirical sizes of tests of model comparisons using known  $\sigma^2$ . Design points are realisations of *bivariate normal* designs with ( $\rho = 0.5$ ). The results listed are the proportion of 500 simulated p-values less than 0.05 under the reduced model. The solid horizontal lines (when shown) indicate a 99% confidence interval from  $\text{Bin}(500, 0.05)$ .

### 4.3.2 Random design: $\sigma^2$ estimated

#### Investigating size

The final scenario to consider within this simulation study, is that of a random design where  $\sigma^2$  is not known and therefore needs to be estimated from the observed data. This scenario is most likely to be encountered in applied settings since rarely, if ever, is the variance of observations around an underlying model known nor are observations usually recorded at design points neatly positioned at regular intervals in the design space. Previous sections have highlighted the complexities associated with random configurations of design points and Chapter 2 considered in detail the estimation of  $\sigma^2$ . Here we encounter these two issues simultaneously.

Given both the importance and the complexities of this setting the simulations will be extended to incorporate a range of smoothing parameters. This allows the construction of ‘size traces’, introduced in Section 1.2.2, which show the empirical sizes of different tests over a range of values for the smoothing parameter. This removes the subjective choice of smoothing parameter from the procedures and gives a more comprehensive description of the tests’ performances.

Table 4.5 lists the conditions and approaches employed in the simulations. Note that only one type of random configuration is listed, namely a uniform distribution on the  $[0, 1]$  plane. Bivariate normal designs, with varying degrees of correlation, will be investigated in Section 4.3.3.

Some of the results of this simulation exercise are shown graphically in Figures 4.17 and 4.18 for sample sizes of 100 and 49 respectively and all are listed in Tables B.12 - B.17 where Table B.11 provides the legend for the different settings. Although the p-values from a CFV comparison of model fits using an uncorrected F distribution are listed in Tables B.16 and B.13, they are omitted from the graphs since they do not offer any improvement over the standard approach. Similarly, the results when the underlying model is linear are also omitted from Figure 4.17 since they always perform well, since the nonparametric fits contain zero bias. Note also, that the approaches incorporating undersmoothed  $(h/2, h/4)$  RSS based estimators of  $\sigma^2$  are not available for all of the values of  $h$  considered. Clearly the fits will deteriorate to interpolations if  $h/2$  or  $h/4$  is too

**Table 4.5.** Summary of the settings and the test procedures used in the simulations of Section 4.3.2.

Data conditions	
design space	uniform-random configuration over $[0, 1]^2$
sample size	49 & 100 design points
error distribution	Normal (zero mean)
error variance	0.01
number of simulations	500
Test approaches	
model comparisons	2d.am vs. 2d.sm 1d.sm vs. 2d.am 1d.sm vs. semi.par.x1 semi.par.x1 vs. 2d.am st.line vs. semi.par.x2
smoothing parameter	0.05, 0.1, 0.15, 0.2, 0.25
variance estimator ( $\sigma^2$ )	RSS ( $h$ , $h/2$ , $h/4$ ) Difference based (unweighted, weighted)
comparisons of model fits	RSSD & CFV
reference distribution	QF, F, corrected F

small.

The behaviour of the tests are best summarised in two groups, reflecting the underlying models:

**Underlying additive model:** this situation poses challenges to all of the test procedures considered. There are no approaches which yield consistent empirical sizes over the range of smoothing parameters considered. Whilst CFV comparisons yield empirical sizes closest to  $\alpha = 0.05$  level, these are very much dependent on the choice of smoothing parameter and the degree of undersmoothing. The most consistent results are obtained via a combination of CFV comparison with a difference based estimator of  $\sigma^2$ , but only at a smoothing parameter of 0.1. There is little to distinguish between a QF and corrected F reference distribution. Interestingly, the standard F test which uses an undersmoothed bandwidth of  $h/2$  does not perform too badly. However, this again emphasises the role the degree of

undersmoothing plays.

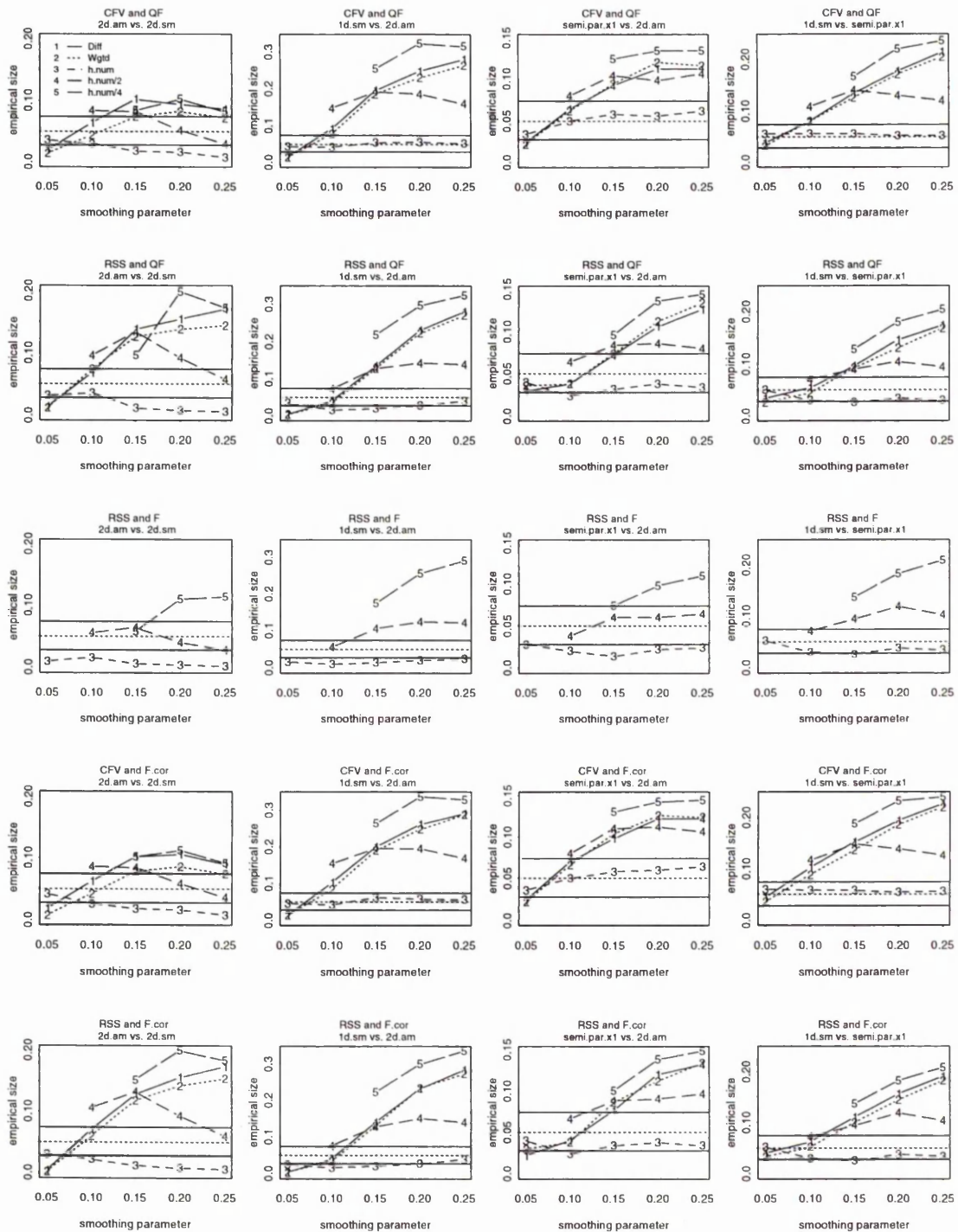
**Other model comparisons :** we note the very favourable behaviour of the CFV statistic with a RSS (not undersmoothed) based estimator of  $\sigma^2$  at all values of smoothing parameter for the remainder of the model comparisons. Although there are values of the smoothing parameter at which other test approaches yield results in line with the specified significance level, these are not consistent across all values of the smoothing parameter considered here. Undersmoothing the fits to obtain  $\hat{\sigma}_{RSS}^2$  can be seen to inflate the size of the tests. Difference based estimators do yield consistent values for lower smoothing parameters, but deteriorate as these increase, reflecting the rising dominance of bias in the CFV and RSSD statistics.

In summary, the results have shown that an approach based on CFV with no undersmoothing to generate  $\hat{\sigma}_{RSS}^2$  returns well calibrated sizes for a range of p-values. The clear advantage over RSSD comparisons is that undersmoothing does not appear to be necessary, thereby removing an unknown 'parameter' from the procedure. The exception to this result is the additive vs. bivariate comparison, where no single test approach yielded consistent empirical sizes.

### Investigating power

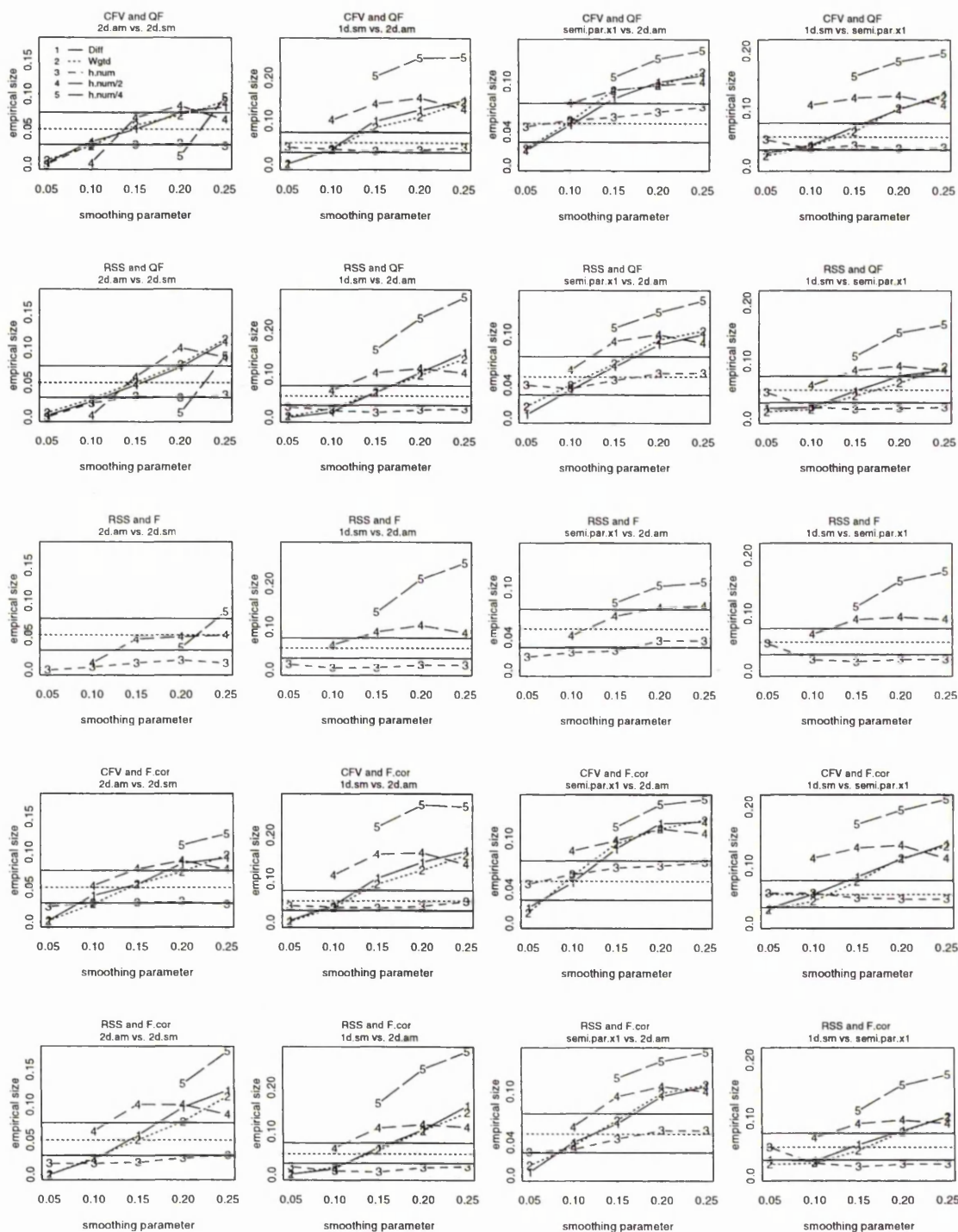
Here we investigate the power performance of the tests considered above which worked well. Given the consistency of the results in all model comparisons except the additive vs. bivariate comparison, the focus is on tests involving a RSS based estimator of  $\sigma^2$  with no undersmoothing over a range of smoothing parameter values. In the case of an additive vs. a bivariate comparison one value of  $h$  is used together with a difference based estimator of  $\sigma^2$ . In both cases a two-moment corrected F reference distribution is used.

The results are listed in Tables B.18 and plotted in Figure 4.19. Once again, the power results do not indicate major differences between test approaches, with a noticeable difference only appearing at very small values of the smoothing parameter. The reduced power of the smaller sample size is very clearly shown.



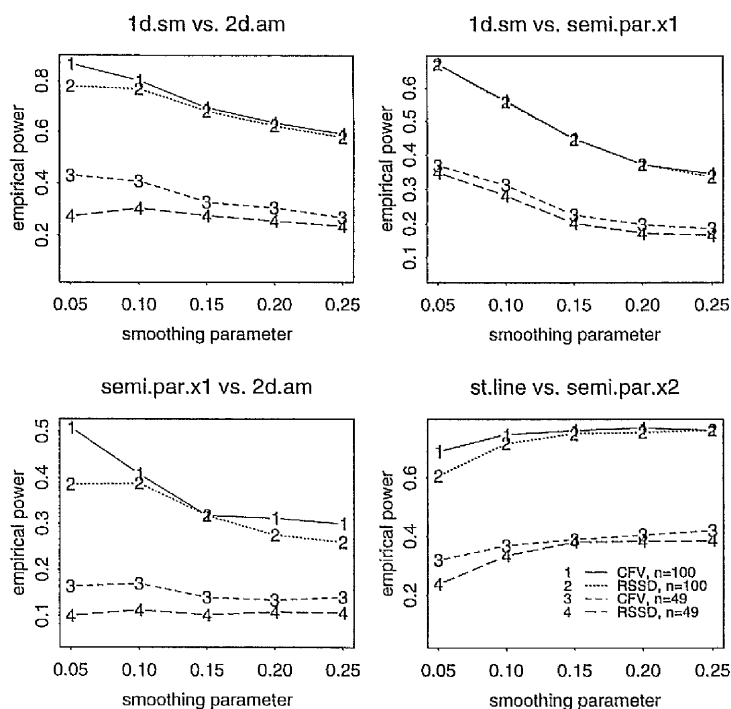
**Figure 4.17.** Size traces showing the performance of model comparisons using 2 comparisons of model fit, 5 estimators of  $\sigma^2$  and 2 reference distributions. Simulations are under the conditions listed in Table 4.5 using a sample size of 100





**Figure 4.18.** Size traces showing the performance of model comparisons using 2 comparisons of model fit, 5 estimators of  $\sigma^2$  and 2 reference distributions. Simulations are under the conditions listed in Table 4.5 using a sample size of 49





**Figure 4.19.** Power results of 500 simulations, over a range of smoothing parameter values, showing the performance of model comparisons using 2 comparisons of model fit, a  $RSS$  based estimator of  $\sigma^2$  (no undersmoothing) and a corrected F reference distributions. Results are shown for four model comparisons, under the same conditions as listed in Table 4.5

### 4.3.3 Random design: correlated covariates

In Sections 3.2.5 & 3.2.6 it was noted that the asymptotic biases of semiparametric and additive model fits contain terms which reflect the joint distribution of the two covariates when they are not independent. In Section 4.3 a finite sample effect of correlation amongst the covariates was also observed. This present section investigates this finite sample behaviour further, attempting to establish at what degree of correlation its effect on the tests' performances is noticeable and what are the best test approaches to guard against the deterioration in the methods of inference.

Although the effect of the dependency amongst covariates is most pronounced when the underlying regression function is of the same form as the model fit, an examination of the expression shows that these dependency terms do not disappear when the true underlying function is of a reduced form. Therefore it is informative and necessary to examine a range of model comparisons under the influence of correlated covariates.

Table 4.6 lists the settings used in the simulations to investigate the size performance of different tests. The settings were chosen in light of the results of Section 4.3.2, reflecting the settings and approaches which were shown to perform well under an uncorrelated random design. Although there was strong evidence that a CFV comparison with a corrected F distribution was the best approach, we retain both the RSSD and (uncorrected) F distribution in the simulation study because this combination tends to be used in practice.

Figure 4.20 displays the results of these simulations. The plots show the effect of increasing the correlation coefficient of the distribution used to generate the covariates. Each panel labels the ten combinations of five smoothing parameters and two comparisons of model fits. The left and right columns correspond to the (uncorrected) F and two-moment corrected F reference distribution respectively.

These plots show clearly how correlation amongst the covariates affects the tests' size performances. As  $\rho$  increases the empirical size of all of the tests increase, with the exception of the univariate vs. semiparametric comparison. This behaviour is expected from the asymptotic bias properties which motivated

**Table 4.6.** Summary of the settings and the test procedures used in the simulations of Section 4.3.3.

Data conditions	
design space	bivariate normal configuration $\mu_1 = \mu_2 = 0.5, \sigma_1 = \sigma_2 = 0.15$ $\rho = 0, 0.2, 0.4, 0.6, 0.8$
sample size	100 design points
error distribution	Normal (zero mean)
error variance	0.01
number of simulations	500
Test approaches	
model comparisons	2d.am vs. 2d.sm semi.par.x1 vs. 2d.am 1d.sm vs. 2d.am 1d.sm vs. semi.par.x1
smoothing parameter	0.05, 0.1, 0.15, 0.2, 0.25
variance estimator ( $\sigma^2$ )	RSS ( $h$ )
comparisons of model fits	RSSD & CFV
reference distribution	F & two-moment corrected F

this section. As the correlation increases, more bias is introduced to the numerator of the test statistic, thus inflating the observed values. This in turn decreases the corresponding p-value thereby inflating the empirical sizes. Changes in the smoothing parameter interact with the correlation effect, yielding a positive effect at the high correlations whilst showing no effect at the low correlations.

The degree to which covariate dependency affects the tests' performance also depends on the model comparison. The most pronounced affect is on the comparison of an additive and bivariate fit, followed by semiparametric vs. additive fits. Both these results are finite sample realisations of the asymptotic results which showed that dependency terms enter these model fits. Although only noticeable at high values of correlation, the comparison of univariate and additive models also shows that these dependency terms remain even when the underlying model is of a univariate form, although the magnitude of the effect of correlation is much less than in the case of an underlying additive regression.

Although deterioration in the performance of the tests is evident, it is also interesting to note under what conditions the tests return results consistent with the  $\alpha = 0.05$  level. Consider the results for additive vs. bivariate comparisons as an illustration. The F reference distribution is clearly conservative relative to its two-moment corrected cousin. It therefore has an advantage over the corrected version when  $\rho$  is in the range 0.3-0.6. In a similar fashion, lower smoothing parameters yield more conservative test results and therefore they are preferable particularly at high levels of correlation. Finally, a comparison of the sizes associated with RSSD and CFV model comparisons reveals that there is much benefit in employing the latter, especially in the presence of moderate covariate correlation. Based on these results, it would seem that inference using a small smoothing parameter, the F reference distribution and a direct CFV model comparison is a sensible approach when moderate to high correlation is present.

Similar guidance can be drawn from the other model comparison results. Note though that the level of correlation at which the test performances deteriorate is approximately 0.4 for semiparametric vs. additive fits and 0.6 for univariate vs. additive.

In summary, these simulations have shown that the presence of correlation amongst the covariates introduces further bias into the small sample properties of test procedures. This property is related to that of 'concurvity', noted by Hastie and Tibshirani (1990) which has an effect, not just on inference but on fundamental issues such as the existence and convergence of solutions to the back-fitting estimator. As such, it is only advisable to use nonparametric modelling when there is little correlation amongst the covariates, in which case the results show that the CFV comparisons with a corrected F distribution work well.

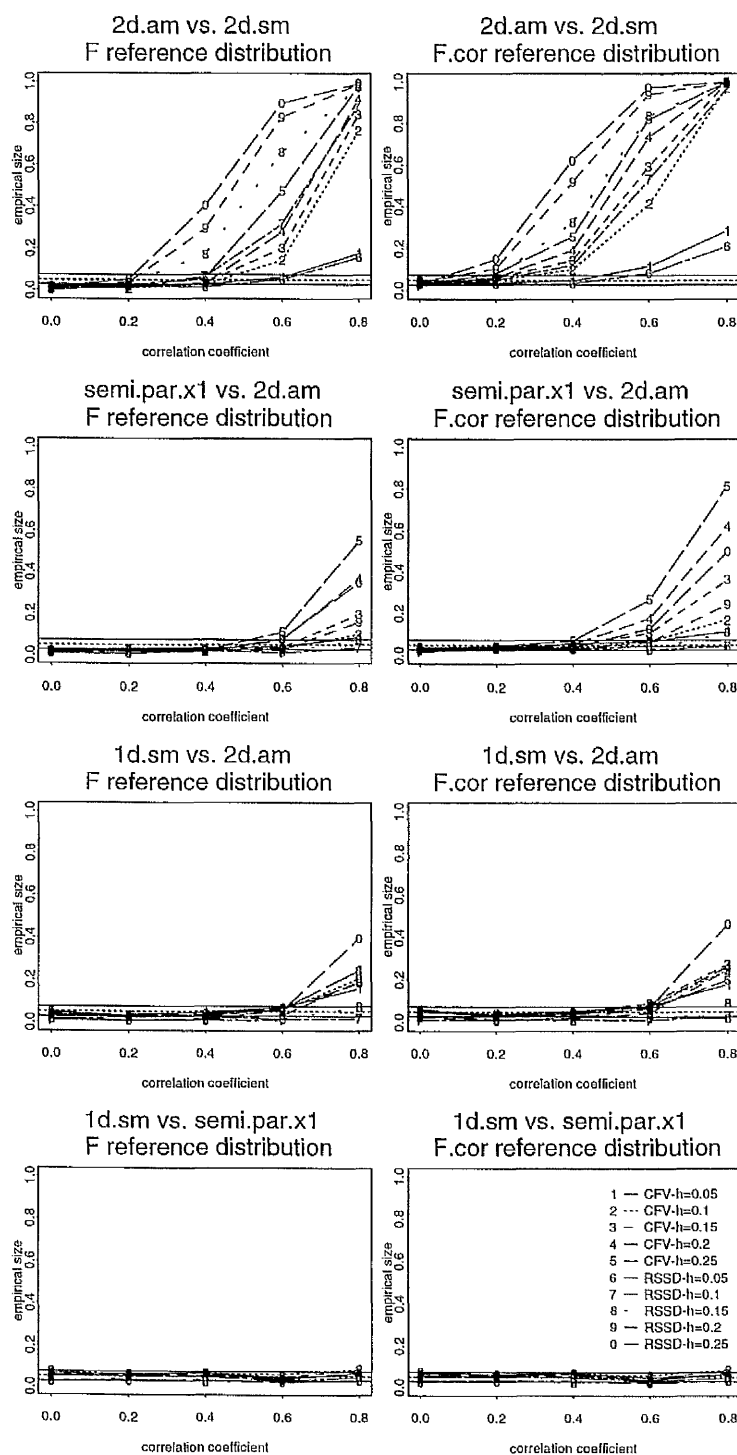


Figure 4.20. Results of simulations described in Table 4.6. The left and right panels correspond to the (uncorrected) F and two-moment corrected F reference distribution respectively.

## 4.4 Model Comparisons Applied to the Reef Data

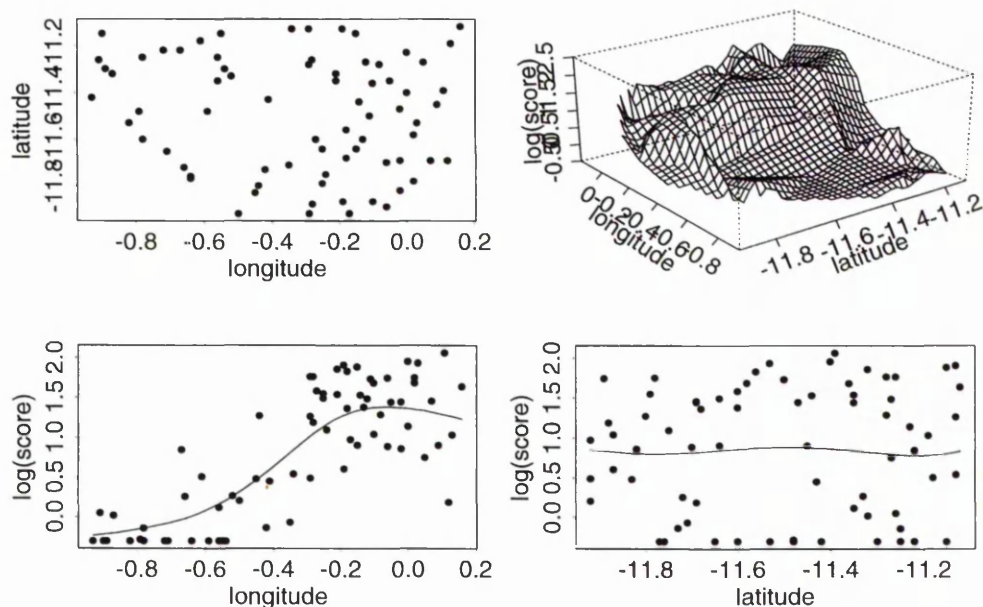
This section presents the results of applying the different approaches to model comparison to a set of real data. The example uses data on the abundance of marine life in the sea bed of the Great Barrier Reef, off the coast of Northern Queensland, Australia. These data were provided by the CSIRO Division of Marine Research, through Dr. Charis Burrridge, and are discussed in Bowman & Azzalini (1997). A selection of the key results of this section are included in Bock (1999).

Figure 4.21 displays data from the CSIRO study for one year only. The survey sampling positions are shown in the top-left panel. The response variable, a catch score, on a log scale, which measures abundance across all species, is shown against the spatial covariates in the remaining plots. The geography of the region allows longitude to be interpreted effectively as distance from the shore, which runs in an approximately North-South direction. Hence, the plots are based on a transformation of the longitude scores, subtracting 143 to simplify the axis labelling, and multiplying by -1 to obtain a west to east perspective.

We will use the tests developed in this chapter to explore the possible covariate effects of latitude and longitude. As a preliminary guide, the univariate plots in the bottom row of Figure 4.21 with superimposed local linear smoothes of the points, a strong effect of longitude is apparent, but there is no immediate evidence of a latitude effect. Since there is clearly an effect of longitude, the question of interest is how does this interact with the latitude direction to determine the mean response of the log of catch score. As such, four models are worth considering:

- ◇ Univariate:  $\log(\text{Score}) = \mu + m_1(\text{longitude}) + \varepsilon$
- ◇ Semiparametric:  $\log(\text{Score}) = \mu + m_1(\text{longitude}) + \beta_2 \cdot \text{latitude} + \varepsilon$
- ◇ Additive:  $\log(\text{Score}) = \mu + m_1(\text{longitude}) + m_2(\text{latitude}) + \varepsilon$
- ◇ Bivariate:  $\log(\text{Score}) = \mu + m_{12}(\text{longitude}, \text{latitude}) + \varepsilon$

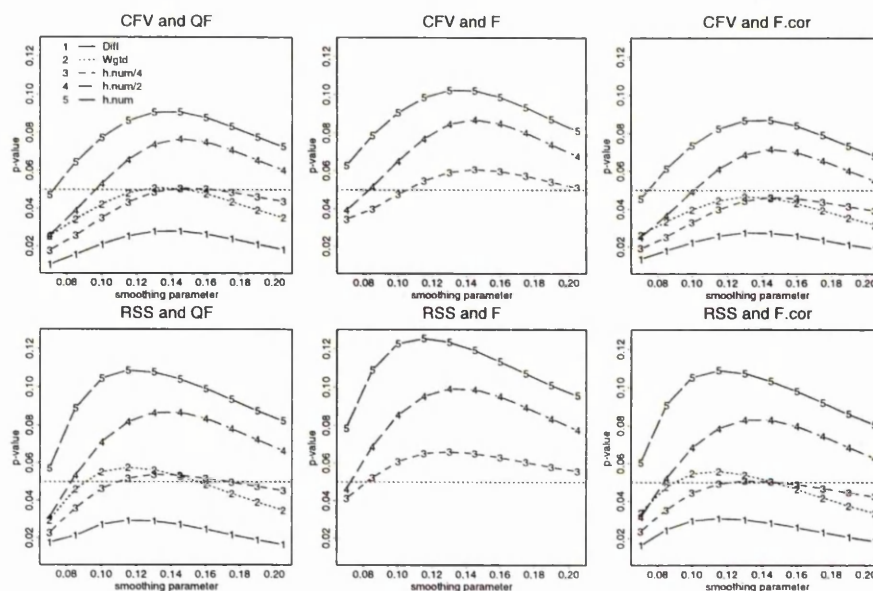
Consider, first, a test to see if there is evidence of a nonlinear effect in latitude. Figure 4.22 displays significance traces for a variety of test approaches



**Figure 4.21.** Plots showing the location of sampling points and the spatial distributions of catch scores from one year of the reef data.

each based on a comparison of an additive fit and a univariate fit to the data shown in Figure 4.21. These results show that different conclusions could be reached depending on the test approach taken. For instance a standard F test, as represented by line '5' in the lower-middle panel of Figure 4.22, consistently returns a nonsignificant p-value at the  $\alpha = 0.05$  level. In contrast, tests using improved estimators of  $\sigma^2$  and corrected reference distributions, return consistently significant results. This highlights the importance of choosing with care a test procedure.

Clearly the sampling locations do not form a regular grid nor do we have knowledge of the magnitude of the error variance of the responses around the hypothesised regression functions. Also, the lack of association between latitude and longitude and the weak correlation of 0.14 observed in the design points, suggest that the covariates are distributed approximately independently. Therefore the results of Section 4.3.2 are a useful guide to which test approach to use. These simulations identified the approach based on a CFV test statistic with an RSS

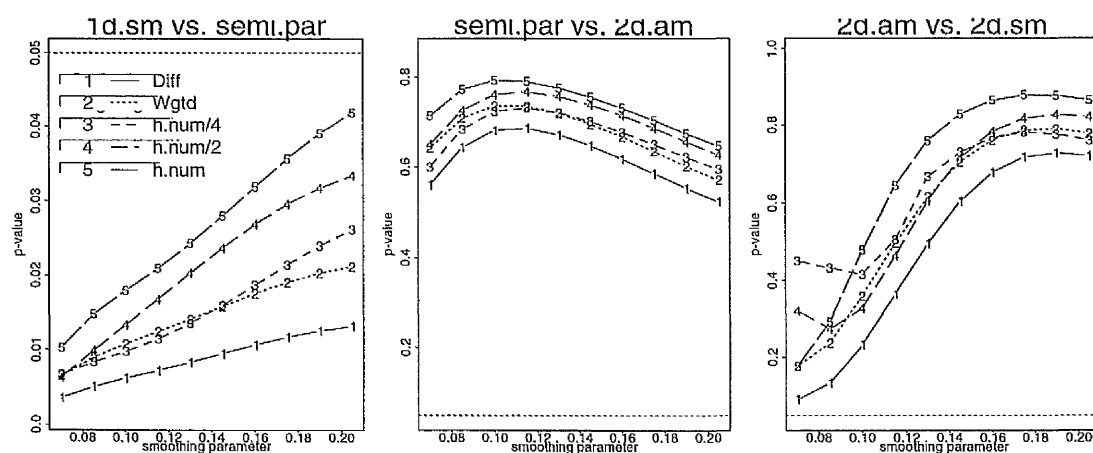


**Figure 4.22.** Significance traces testing a univariate smooth fit in longitude against an additive fit for each combinations of model comparison, variance estimator and reference distribution.

based estimator of  $\sigma^2$  and a corrected F distribution as the most appropriate test procedure. This corresponds to line '5' in the upper right panel of Figure 4.22. As such, we conclude that there is insufficient evidence that the extra variability captured by the smooth latitude term reflects a nonlinear effect in latitude in the underlying regression surface. Note that this is true over the range of smoothing parameters considered.

This conclusion leads to the consideration of other possible model fits, firstly a univariate smooth versus a semiparametric fit, as shown in the first panel of Figure 4.23. Only 1 test procedure, that using a CFV statistic and a corrected F reference distribution, is presented since it was shown to be most favourable in the simulation study. In this case the results using all five estimators of  $\sigma^2$  are conclusive: there is evidence of a linear effect in latitude. The remaining panels of Figure 4.23 confirm this result by considering, once again, more general fits, i.e. additive and semiparametric. In these cases the conclusions are also unanimous across the different estimators of  $\sigma^2$ . Therefore, at the  $\alpha = 0.05$  level we clearly conclude that there is evidence of a nonparametric effect in longitude and a linear





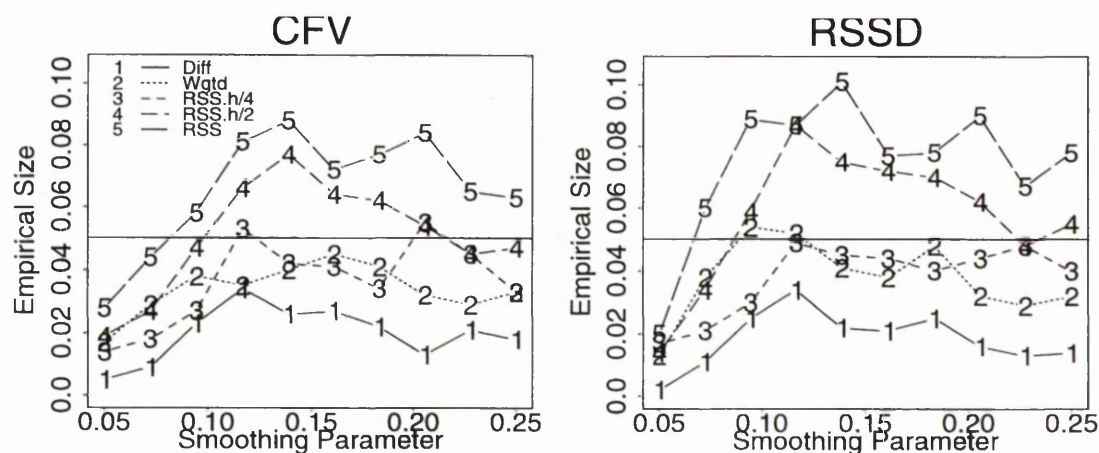
**Figure 4.23.** Significance traces of tests based on CFV and corrected F distribution for comparisons amongst semiparametric, additive and bivariate smooth fits to the reef data.

effect in latitude.

Bootstrap techniques are often proposed as ‘solutions’ to problems of inference where standard assumptions no longer hold. In the present case, however, there is more than a question of a suitable reference distribution to answer. The choice of CFV and RSSD model comparisons remain as does the question of what estimate of  $\sigma^2$  to use. However, the bootstrap approach has been suggested by several referees and colleagues and therefore it will be presented here as an illustration.

Given that most uncertainty surrounds the comparison of a univariate smooth fit and an additive fit, this comparison will be the focus of the bootstrap simulations. They proceed as follows:

1. Fit the univariate smooth model to the data
2. Fit the additive model to the data
3. Estimate  $\sigma^2$
4. Calculate the test statistic
5. Generate  $n$  simulated ‘errors’ from  $N(0, \hat{\sigma}^2)$
6. Add the simulated errors to the fitted model in 1.



**Figure 4.24.** Bootstrap p-values for the 1d.sm vs. 2d.am model comparison applied to the reef data.

7. Fit a univariate smooth and an additive model to the simulated data in 6.
8. Calculate test statistic from fits in 7.
9. Repeat steps 5.- 8. 1000 times
10. Calculate the bootstrap p-value as the proportion of the 1000 simulated test statistics exceeding the original test statistic calculated in 4. from the original data.

Clearly, there are a number of open questions remaining in the procedure above: smoothing parameters are required in Steps 1,2 and 7; different estimators of  $\sigma^2$  could be used in Step 3; test statistics in Steps 4. and 8. could use either CFV or RSSD comparisons of model fits. As in the simulation exercises, combinations of these settings were investigated using the bootstrap approach.

The results of these bootstrap simulations are shown in Figure 4.24. These correspond to jagged versions of the last column of Figure 4.22 which confirms that the corrected F reference is a good approximation to the unknown reference distribution. Apart from confirming this aspect of the test procedure, however, the bootstrap approach fails to guide us to which test approach leads to the 'correct' conclusion.

Before leaving the reef data it is instructive to compare the results of our methods with those obtained via standard S-Plus functions which fit and compare nonparametric models. The `gam()` function includes, in its scope of model fits, the bivariate class of models. It permits the specification of two types of smoothers, splines via the `s()` argument and nearest neighbour local linear smoothers via `lo()` (short for Cleveland and Devlin's (1988) `loess`). Because of the similarities with the local linear approach, `loess` smoothers were used to fit and compare different nonparametric models relating longitude and latitude to the response variable.<sup>4</sup>

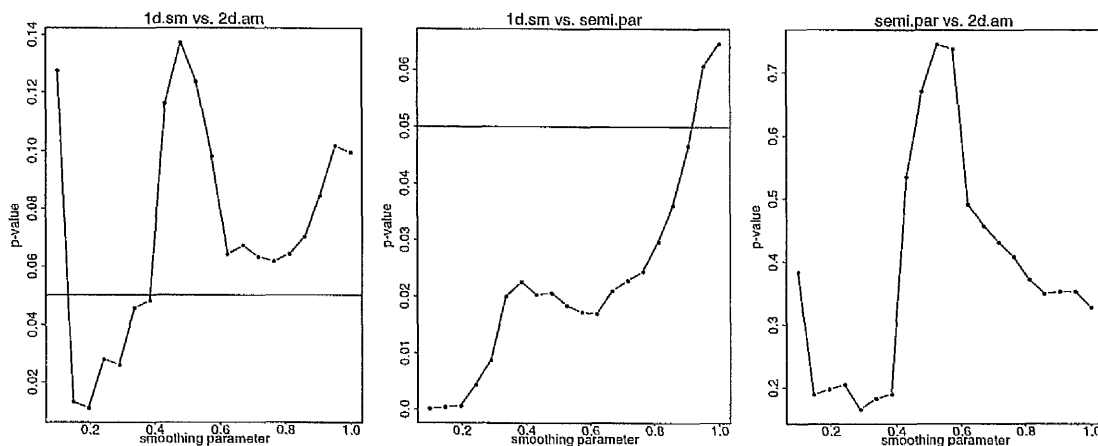
Figure 4.25 shows significance traces for three `gam()` model fits to the reef data. The first panel shows a comparison between a univariate nonparametric model in longitude and an additive model in both latitude and longitude. The p-values are inconclusive falling, as they do, both beneath and above the  $\alpha = 0.05$  significance level over the range of smoothing parameters considered. The second panel compares the same univariate fit with a semiparametric model permitting latitude to appear linearly. Here the conclusion is as before, i.e. there is evidence of an effect in latitude. The third panel compares this semiparametric fit with an additive fit, thereby confirming the conclusions previously reached, i.e. a linear effect in latitude appears sufficient.

Although the conclusions of the two approaches are the same, a comparison of Figures 4.22, 4.23 and 4.25 reveals some interesting differences. Firstly, the comparison of a univariate and an additive fit using the S-Plus functions was inconclusive whereas the test suggested by the simulation results clearly failed to reject the univariate fit. This is (in part) due to the different approaches to model comparisons, since the S-Plus p-values are derived from approximate F tests. A suitable comparison, therefore, is between the first panel of Figure 4.25 and line '5' of the lower-middle panel of Figure 4.22. Even here there are marked differences between the two approaches.

These differences are because the approaches differ not only with respect to the way they compare model fits but also in the way they construct the fits. For instance, the significance traces of Figure 4.25 highlight the different form of the smoothing parameter used in the `lo()` fit. The nearest neighbour definition

---

<sup>4</sup>The smoothing spline fits were also explored and the results (not shown here) are very similar to those presented here.



**Figure 4.25.** Significance traces of tests using the `gam()` and `lo()` functions in S-Plus for comparisons amongst univariate, semiparametric and additive model fits to the reef data.

of 'local' specifies a certain proportion ('span') of the data to be included in the local model fit. A consequence is that significance traces no longer vary smoothly as the smoothing parameter changes but rather are prone to sudden changes as seen in each panel of Figure 4.25. This is an undesirable quality since ideally the choice of smoothing parameter should play a minor role in matters of inference. Sudden changes in p-values relative to changes in the smoothing parameter do not comply with this ideal.

Another serious consequence of this form of the smoothing parameter is that the *nested* properties of competing models are no longer guaranteed. In this current example, for instance, when S-Plus fitted both an additive and a bivariate smooth model to the data using the same smoothing parameter ('span') in both, the RSS of the bivariate model were in fact *greater* than that of the additive fit. This clearly makes comparisons based on RSSD nonsensical since the F distribution is strictly nonnegative. The situation is made even more absurd by S-Plus' 'solution' to this difficulty, that is using the *absolute* value of RSSD in the test statistic, which renders the subsequent p-value misleading as well as meaningless. This observation is an argument against the use of nearest neighbour definitions of local-ness and indeed favours the use of the same smoothing parameter in the local linear regression comparisons. This and other advantages of the approaches

developed in Chapter 3 and implemented in this chapter are discussed in the following sections.

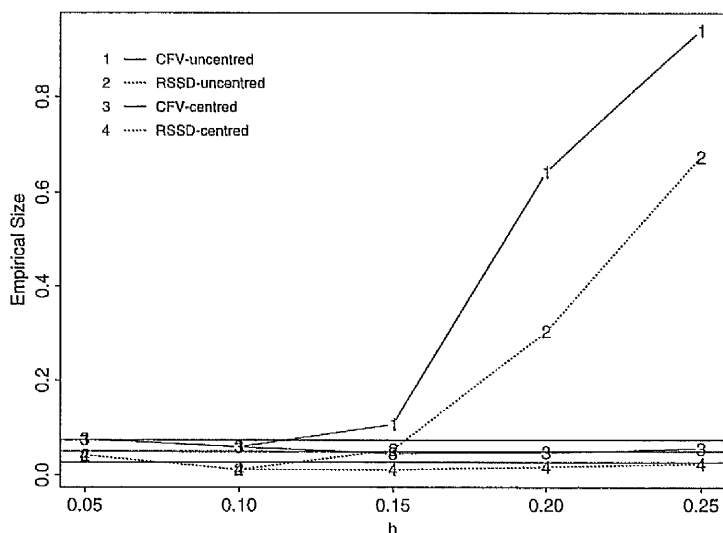
## 4.5 Discussion

Chapters 3 and 4 have described and investigated several approaches to non-parametric model inference via comparisons of model fits. These approaches have included standard tests analogous to the linear parametric setting and new approaches inspired by the properties of the local linear regression estimates of the regression function.

Before considering the different approaches investigated in the simulations, it's worthwhile highlighting several characteristics which differ from the standard practice. Two aspects of all of the approaches considered have been the use of *centred smoothers* and the *same smoothing parameters* in the model fits. Both of these were motivated by the asymptotic bias results which indicated that the biases were more likely to cancel, especially in the CFV statistic, under these conditions.

To illustrate the improvement offered by these approaches in a finite sample setting, several simulations were conducted. In each case 500 data sets were simulated over uniform-random designs with 100 points using the univariate regression function used throughout the simulations and an error variance of 0.01. To this simulated data were fitted two models: a univariate smooth model and a bivariate additive model. Given the results of Section 4.3.2 tests using a RSS based estimator of  $\sigma^2$  with no undersmoothing and a corrected F reference distribution were used. Both the CFV and RSSD comparison of model fits were kept to highlight their differences. In each case, the empirical size at the  $\alpha = 0.05$  level is recorded at several values of the smoothing parameter values.

In the first instance, size results from a comparison involving a univariate smooth fit using a centred smoother were compared with equivalent results using an *uncentred* univariate smoother. A review of the asymptotic bias expressions in Section 3.2.4 reveals that this will introduce a difference in the asymptotic biases of  $\frac{h^2}{2}\mu_2(K)E(m''(\cdot))$ . Figure 4.26 shows that the finite sample behaviour reflects this asymptotic result. For small values of  $h$  there is little to distinguish between

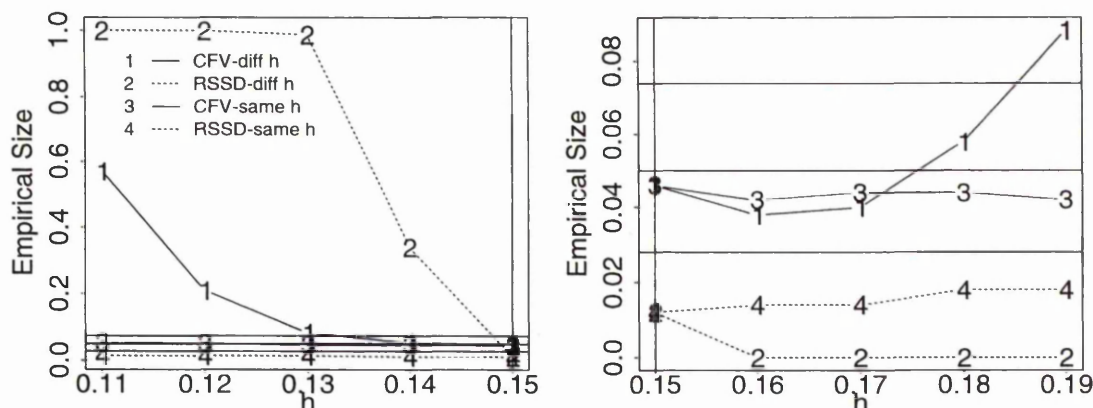


**Figure 4.26.** Size traces of four tests to compare a univariate smooth and a bivariate additive fit. Tests ‘3’ and ‘4’ use centred smoothers throughout, tests ‘1’ and ‘2’ use uncentred smoothers in the univariate smooth fit.

the two approaches and the empirical sizes are equivalent to the corresponding simulations in Section 4.3.2. As  $h$  increases beyond 0.15, however, the sizes of both CFV and RSSD model comparisons deteriorate rapidly, returning an increasing proportion of type I errors.

In practice, however, it would be most unlikely for a centred univariate smoother to be employed since this requires modifying the standard output of functions such as those in S-Plus. Indeed, the results presented in Figure 4.25 are not based on centred smoothers but rather the default fits given by S-Plus. Therefore it is worth emphasising this aspect of our approaches to model comparisons.

The second consideration is the effect of using a different smoothing parameter in each of the model fits. Since the bias of the fits is a function of the smoothing parameter, differences in the smoothing parameters will yield differences in the biases which in turn will impact the performances of the tests. Figure 4.27 shows results obtained when the smoothing parameter used to fit the univariate smooth fit is held at 0.15 but the smoothing parameters used in the two dimensional



**Figure 4.27.** Size traces of four tests to compare a univariate smooth and a bivariate additive fit. Tests '3' and '4' use equivalent smoothing parameters in the two fits. Tests '1' and '2' use a range of smoothing parameters in the additive fit but keep the smoothing parameters in the univariate smooth fit constant at 0.15.

additive fit vary between 0.11 and 0.19.

In both the CFV and RSSD model comparisons there is a very marked effect of using different smoothing parameters, although the pattern in each is quite different. This difference can be understood by considering the expressions for the CFV and RSSD statistics presented in Equations 4.3 and 4.4 respectively.

In the CFV statistic the biases of the two fits appear in the form  $(\mathbf{b}_R - \mathbf{b}_F)^T(\mathbf{b}_R - \mathbf{b}_F)$ . Hence any difference in the model biases will always appear as a positive term in the CFV statistic, thereby inflating its value, and hence reducing the associated p-value. This is why departures either side of the null model smoothing parameter induce increased sizes.

For the RSSD statistic the bias terms appear as  $\mathbf{b}_R^T \mathbf{b}_R - \mathbf{b}_F^T \mathbf{b}_F$  and therefore the relative size of the bias terms determines the direction of the impact on the value of the test statistic. Smaller values of smoothing parameters in the full model, relative to those used to fit the null model, will cause an increase in the RSSD statistic and vice versa. This explains the shape of the RSSD test lines in Figure 4.27.

In practice, the use of automatic smoothing parameter selection methods may

lead to the use of different smoothing parameters. These choices may indeed optimise certain properties of the fits for the purposes of estimation but this illustration has demonstrated that, for purposes of inference, equivalent smoothing parameters are preferable.

An interesting consequence of this behaviour is that it calls into question methods of smoothing parameter selection based on model comparisons. Authors such as Cleveland & Devlin (1988) and Hastie & Tibshirani (1990) suggest the use of model comparisons of the type considered here but between two model fits of the same class (eg. both univariate smooths) each using a different value of the smoothing parameter(s). The result of the test is used to determine whether the increased complexity of the model with the smaller smoothing parameter is warranted by its improvement to the fit of the data. However, this illustration has shown that care must be taken in this setting since the properties of the tests may be corrupted by the differences in the biases of the model fits.

The importance of using centred smoothers and equivalent smoothing parameters has therefore been illustrated. These approaches were used throughout the simulations, although there is nothing in the research literature which emphasises this approach to model comparisons.

The simulations themselves have brought many important factors regarding the merits of different approaches to model comparisons to light. There is evidence, for instance, that a CFV test statistic is a sensible approach to model comparison, offering substantial improvements in the consistency of the empirical sizes in some settings and never returning significantly worse results in others. The two-moment corrected F distribution was also observed to work well, significantly better than the uncorrected F and at least as well as the distribution derived from the properties of a quadratic form. The appropriateness of this distribution was further confirmed by the bootstrap simulations applied to the reef data.

One of the most significant insights these results have provided is into the consequences of different designs on the test procedures. In the case of regular grids, it was noted that the asymptotic results concerning the bias of local linear fits hold for small samples. This accounts for the simulation results which showed that the use of an accurate estimate of  $\sigma^2$  was of supreme importance. This was



borne out by both the results using the exact (true) value of  $\sigma^2$  (Section 4.2.1) and the results using a number of different estimators of  $\sigma^2$  (Section 4.2.2). In the former case as long as a 'corrected' reference distribution was consulted the size results were consistent. In the latter case, the difference based estimator of  $\sigma^2$  was the preferred approach, followed closely by an undersmoothed RSS based estimator. Therefore the recommended approach to model comparisons when the data is observed over a regular grid is to use a CFV test statistic with a difference based estimator of  $\sigma^2$  and a two moment corrected F reference distribution.

In the case of the random design, marked differences were observed from the regular grid setting, due to the extra presence of finite sample bias in the test statistics. The key factor remained the estimation of  $\sigma^2$ , however the use of the most accurate estimator did not guarantee a good test performance. Indeed, in the simulations where the true value was used (Section 4.3.1) the empirical sizes were inconsistent for most of the settings used and indeed were inappropriate at all settings for certain model comparisons.

Fortunately, in the more realistic scenario of using an estimate of  $\sigma^2$  in the model comparisons, some approaches returned consistent results. Most interestingly the 'small bias', i.e. difference based and undersmoothed RSS based, estimators of  $\sigma^2$  did *not* yield the best results. Indeed the best approach in nearly all comparisons was a CFV test statistic using an RSS based estimator of  $\sigma^2$  from the full model fit used to define CFV and a corrected F distribution. This was true over a range of smoothing parameters.

This can be understood, once again, by considering the finite sample bias which enters the test statistic. It was noted in Section 4.3 that over random designs, the finite sample biases of the model fits do not cancel exactly. Therefore bias is present in the numerator and this increases with the smoothing parameter. Minimising the bias in the denominator by estimating  $\sigma^2$  (using difference based methods or using the RSS from undersmoothing the full model fit) does not therefore enhance the test statistics' distributional properties. Rather, the results indicate that there is a 'balancing' effect between the bias of the CFV statistic and the bias of the standard RSS  $\sigma^2$  estimator which yields a test with an accessible reference distribution when the null hypothesis is true.

Lastly, the simulations have shown that in the presence of low to moderate

correlation amongst the covariates, the results from the independent covariate setting can be used to guide the approach to model inference. The exception to this is a comparison of the most general models: a bivariate additive and a bivariate smooth. Comparisons amongst such fits are very sensitive to a lack of independence amongst the covariates. Even so, tests using an undersmoothed estimate of  $\sigma^2$  returned consistent results even at a reasonably high level of correlation.

In summary, the simulations reported in this chapter have shown the usefulness of tests based on a statistic which compares and measures discrepancies of model fits directly. It has also been seen that the appropriate choice of an estimator of  $\sigma^2$  is dependent on the distribution of the design space. Minimum bias estimators are favoured in the regular grid setting, whereas estimators based on the fit used in the model comparison perform better with random designs. A two-moment corrected F distribution was found to perform well and owing to its accessibility it is the recommended reference distribution. Chapter 5 will build on these findings by extending these methods to models with an unlimited number of covariates.

## Chapter 5

# Inference Beyond Two Covariates: Semiparametric Additive Models

### 5.1 Introduction: Models Under Consideration

The previous chapter focused on methods of inference amongst a restricted class of models. These involved at most two covariates, appearing either as linear or smooth terms. In this chapter we shall examine inference amongst a class of models which is a natural extension beyond two covariates.

All of the models considered in this chapter are semiparametric in that they comprise two components: one linear the other nonparametric. The linear component of the models may involve any number of terms. The smooth component, however, is limited to a maximum of two dimensions, appearing either additively or in a single bivariate smooth function. In this way, they extend the models of the previous chapter by allowing for any number of linear terms.

The terminology ‘Semiparametric Additive Models’ (SAMs) was introduced by Opsomer and Ruppert (1999) to describe models with an unlimited number of linear and univariate smooth terms combining additively. This terminology is also used here, although with the important distinction that at most two predictors appear nonparametrically. Models which belong to our SAM class have one of

the following forms:

### Linear

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (5.1)$$

### Univariate-SAM

$$\mathbf{y} = m_1(\mathbf{x1}) + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (5.2)$$

### Additive-SAM

$$\mathbf{y} = m_1(\mathbf{x1}) + m_2(\mathbf{x2}) + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (5.3)$$

### Bivariate-SAM

$$\mathbf{y} = m(\mathbf{x1}, \mathbf{x2}) + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (5.4)$$

where  $\mathbf{y}$  is an  $n$ -vector of responses,  $\mathbf{x1}$  &  $\mathbf{x2}$  are  $n$ -vectors of covariates,  $\mathbf{Z}$  is an  $n \times (p+1)$  matrix consisting of one column of 1's and  $p$  columns of covariate  $n$ -vectors  $(\mathbf{z1}, \dots, \mathbf{zp})$ ,  $\boldsymbol{\beta}$  is a  $(p+1)$ -vector of unknown constants  $\beta_0, \beta_1, \dots, \beta_p$ ,  $\boldsymbol{\varepsilon}$  is an  $n$ -vector of  $N(0, \sigma^2)$  error terms. The functions  $m$  represent nonparametric components, where  $m_1(\cdot)$  and  $m_2(\cdot)$  are smooth functions in one dimension and  $m(\cdot, \cdot)$  defines a two dimensional smooth surface. Note that the 'prefix' in the terminology refers to the nonparametric component<sup>1</sup>.

The motivation for considering such models is that they offer extra versatility yet retain attractive properties useful for inference. The unlimited number of linear terms is an attractive feature which extends the scope of their applicability to many practical problems. For instance, multiple regression may concentrate on a set of covariates in the presence of one or two 'nuisance' variables, the form of whose influence on the response is not directly relevant. These models allow for the inclusion of such variables as nonparametric components in three forms (univariate, additive, bivariate), which progressively admit greater flexibility, thereby potentially yielding a parsimonious model.

In addition to their applicability, these models have the attractive property that explicit definitions of the model fits exist in each case. As in Chapter 3,

---

<sup>1</sup>The 'Univariate SAM' is more commonly known as a *partial linear model* in the literature, however, the 'SAM' terminology highlights the differences between the forms considered here.

the backfitting algorithm can be used to estimate the unknown parameters and functions using simple least squares and local linear smoothers. Hence, expressions for the asymptotic bias of these fits can be derived. The forms of these expressions suggest ways in which methods of inference can proceed.

It is possible to derive recursive expressions for the estimates from the general class of SAMs (unlimited number of nonparametric components), as described by Opsomer (1999). However, direct computation of the model fits using these expressions is much less efficient than via the iterative scheme of the backfitting algorithm. Furthermore, asymptotic bias expressions could be derived from these, but since these too would be recursive and indeed quite complicated, we shall restrict attention to the class defined above.

Before proceeding, however, it is worth noting other generalisations which combine both linear and nonparametric terms. *Single index models* combine these two components by fitting a nonparametric curve using a linear combination of all the covariates. That is, the dimensionality problem is overcome by reducing the  $p$  covariates to a single predictor. This requires the estimation of  $p$  coefficients to perform the linear transformation and an estimate of the nonparametric curve. Härdle, Hall and Ichimura (1993) consider the use of the Nadaraya-Watson estimator (see Section 1.2.1) but any smoother could be used.

An extension of this model is the *partially linear single index model* which allows covariates to also enter linearly outside the nonparametric term. Carroll *et al.* (1997) generalized this model by allowing for non-normal error distributions and non-identity link functions. Hence, 'Generalized Partially Linear Single Index Models' (GPLSIMS) relate the response variable  $Y$  and multidimensional covariates  $X$  and  $Z$  by

$$g\{E(Y|X = \mathbf{x}, Z = \mathbf{z})\} = m(\boldsymbol{\alpha}^T \mathbf{x}) + \boldsymbol{\beta}^T \mathbf{z}$$

where  $g$  is the known *link* function,  $m(\cdot)$  is an unknown smooth univariate function, and  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  are unknown parameter vectors. The distribution of the responses (with conditional means defined by the model above) can be any belonging to the exponential family of distributions. The authors use maximum quasi-likelihood (McCullagh and Nelder, 1989) together with local linear regression to

estimate the unknown quantities.

Severini & Staniswalis (1994) describe the properties of fits from a class of semiparametric models which includes the bivariate-SAM. The class is similar in form and generality to GPLSIMs the difference being that the single nonparametric term is a *surface smoother*. Their approach is based on the concept of ‘generalized profile likelihood’ (Severini and Wong (1992)) and they employ kernel (local constant) regression to estimate the nonparametric component. They note the undesirable boundary bias properties of this method and suggest local linear regression as a solution, but instead choose a trimming algorithm which essentially only uses interior design points to fit the model.

Essentially, SAMs differ from each of these in that they allow more than one nonparametric term to enter the model. In terms of their ‘coverage’ they lie between GPLSIMs and Quasi-likelihood Semiparametric Models of Severini and Staniswalis (1994). We do not consider ‘generalisations’ (in the GLM sense) here since, as noted earlier, this requires iterative fitting which complicates the properties of the estimates. SAMs are favoured here because they occupy a ‘middle ground’ and because their properties are conducive to inference, as the following section will show.

## 5.2 Defining the Fitted Values and Bias

Chapter 3 showed that each of the models in the bivariate class considered there had explicit solutions to the backfitting algorithm and therefore did not require iterative methods for estimation. In this chapter we shall extend this result by defining explicitly the fitted values of each form of the the SAMs introduced in Section 5.1. These results exist for any number of linear terms included in the model.

In order to define the fits of each model explicitly, consider a general form of the models under investigation in this chapter:

$$y = M(\mathbf{x1}, \mathbf{x2}) + \mathbf{Z}\boldsymbol{\beta} + \varepsilon ,$$

where  $M(\mathbf{x1}, \mathbf{x2})$  is an  $n$ -vector of one of the three nonparametric components

described earlier, evaluated at the design points. That is,

$$\begin{array}{ll} \text{Univariate-SAM} & M(\mathbf{x1}, \mathbf{x2}) = m_1(\mathbf{x1}) \\ \text{Additive-SAM} & M(\mathbf{x1}, \mathbf{x2}) = m_1(\mathbf{x1}) + m_2(\mathbf{x2}) \\ \text{Bivariate-SAM} & M(\mathbf{x1}, \mathbf{x2}) = m(\mathbf{x1}, \mathbf{x2}) \end{array}$$

Consider, first, the task of obtaining estimates of the linear parameters contained in  $\beta$ . If the nonparametric component was not present, the least squares fit to the data would return fitted values  $\hat{\mathbf{y}} = \mathbf{Z}\hat{\beta} = \mathbf{H}\mathbf{y} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}$ . Thus, the linear component has associated with it a *projection* matrix,  $\mathbf{H} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$ , which is just a special type of a *smoothing* matrix.

A more familiar type of smoothing matrix,  $\mathbf{S}$  say, is associated with the nonparametric component. Section 3.2 defined the smoothing matrices corresponding to a local linear fit for each of the nonparametric components described here. If the linear component was not present, then the fitted values of the nonparametric component using a linear smoother would be given by  $\hat{M}(\mathbf{x1}, \mathbf{x2}) = \mathbf{S}\mathbf{y}$ . Although we will continue to employ local linear smoothing the following expressions for model fits hold for any linear smoother.

Clearly though, both components are present and need to be estimated simultaneously. Several estimators have been proposed in the setting of a single univariate nonparametric term, sometimes called *partial linear models* in the literature. Initial approaches, see Engle *et al.* (1986) and Shiau *et al.* (1986), used penalised least squares to estimate both components. The term ‘partial smoothing spline’, introduced by Wahba (1984) to describe the estimate of the nonparametric component, reflects the cubic smoothing spline solution to one form of this approach. Green *et al.* (1985) suggest replacing the smoothing matrix equivalent to spline smoothing in Wahba’s treatment with an arbitrary linear scatterplot smoother, our  $\mathbf{S}$  introduced above.

It can be shown, through a parameterisation of the nonparametric component (Speckman 1988, Section 3), that one set of defining equations is

$$\hat{\beta} = (\mathbf{Z}^T(\mathbf{I} - \mathbf{S})\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{I} - \mathbf{S})\mathbf{y} \quad (5.5)$$

$$\hat{\mathbf{m}} = \mathbf{S}(\mathbf{y} - \mathbf{Z}\hat{\beta}). \quad (5.6)$$

Combining these two estimators, it is clear that the smoothing matrix  $\mathbf{S}_{SP}$  which yields the fitted values of the *entire* model (not just its components), i.e.  $\hat{\mathbf{y}} = \mathbf{S}_{SP}\mathbf{y}$  is

$$\mathbf{S}_{SP} = \mathbf{S}(\mathbf{I} - \mathbf{Z}(\mathbf{Z}^T(\mathbf{I} - \mathbf{S})\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{I} - \mathbf{S})) + (\mathbf{Z}^T(\mathbf{I} - \mathbf{S})\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{I} - \mathbf{S}). \quad (5.7)$$

It turns out that these estimators correspond to solutions to the *backfitting algorithm*. To see this, recall that in the semiparametric context, backfitting proceeds by iteratively smoothing and updating the *partial residuals* relative to the two components, in the following way:

**initialise**  $\hat{\boldsymbol{\beta}} = (\bar{y}, \beta_1^{(0)}, \dots, \beta_p^{(0)})^T$

**estimate** the nonparametric component  $\hat{M}(\mathbf{x1}, \mathbf{x2}) = \mathbf{S}(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}})$

**estimate** the parametric component  $\hat{\boldsymbol{\beta}} = \mathbf{H}(\mathbf{y} - \hat{M}(\mathbf{x1}, \mathbf{x2})) = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{y} - \hat{M}(\mathbf{x1}, \mathbf{x2}))$

**repeat** steps 2. and 3. until the estimates don't change

Since there are only two steps which are repeated, explicit solutions for  $\hat{M}(\mathbf{x1}, \mathbf{x2})$  and  $\hat{\boldsymbol{\beta}}$  exist, namely 5.5 and 5.6 above. These estimators taken together will henceforth be known as the *backfitting estimator*.

Often when these results are presented, as in Hastie & Tibshirani (1990, pp. 118), it is assumed implicitly that matrix  $\mathbf{S}$ , and therefore the nonparametric component, consists of a single univariate nonparametric term. However, as Hastie & Tibshirani (1990, pp. 115) note and Speckman's (1988) notation makes clear, the result is general, allowing for different types of smoothers (although they must be linear) and indeed smoothers based on any number of covariates. Since the three forms of  $M(\cdot, \cdot)$  each have an explicit fit, they also have an associated smoothing matrix. Substituting these matrices into Equation 5.7 yields explicit fits for each of the models under investigation in this chapter.

It is worth noting here that alternate forms of estimators for partial linear models have been suggested. Prompted by results of Rice (1986) highlighting



an unattractive bias property of  $\hat{\beta}$  above when correlated covariates are used, Speckman (1988) proposed an alternate method motivated by partial regression plots. The approach is to form partial residual vectors adjusting for  $\mathbf{x1}$ , say, by defining  $\tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{S})\mathbf{y}$  and  $\tilde{\mathbf{Z}} = (\mathbf{I} - \mathbf{S})\mathbf{Z}$ . By analogy with ordinary least squares, the estimators become:

$$\hat{\beta}_{SPK} = (\tilde{\mathbf{Z}}^T \tilde{\mathbf{Z}})^{-1} \tilde{\mathbf{Z}}^T \tilde{\mathbf{y}} = (\mathbf{Z}^T (\mathbf{I} - \mathbf{S})^T (\mathbf{I} - \mathbf{S}) \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{I} - \mathbf{S})^T (\mathbf{I} - \mathbf{S}) \mathbf{y} \quad (5.8)$$

$$\hat{\mathbf{m}}_{SPK} = \mathbf{S}(\mathbf{Y} - \mathbf{Z} \hat{\beta}_{SPK}). \quad (5.9)$$

Speckman (1988) identifies two instances where the estimators 5.8 and 5.9 perform better (in terms of asymptotic bias of  $\beta$ ) than estimators 5.5 and 5.6. The first is where the nonparametric covariates  $X$  and the parametric covariates  $Z$  are correlated. The second is when the smoothing parameter is chosen via cross-validation.

In defence of the use of Equations 5.5 and 5.6, however, Opsomer and Ruppert (1999) considered the backfitting estimator in the specific context of local linear regression. They proposed an alternate method of smoothing parameter selection which yielded parameter estimates with properties equivalent to those via Speckman's methods. Given this result together with the property that the backfitting paradigm extends easily to more than one nonparametric term, leads us to only consider the backfitting estimator in this chapter. The effect of correlated covariates is reflected clearly in the expressions for asymptotic bias presented in Sections 5.2.1- 5.2.3.

Given explicit definitions, properties such as the finite sample bias and covariance matrix of the estimators follow directly, namely

$$\text{Bias}(\hat{\mathbf{y}}) = \mathbf{S}_{SP} M(\mathbf{x1}, \mathbf{x2}) - M(\mathbf{x1}, \mathbf{x2})$$

$$\text{cov}(\hat{\mathbf{y}}) = \sigma^2 \mathbf{S}_{SP} \mathbf{S}_{SP}^T.$$

From these definitions, the bias can be expressed in terms of the two component biases, labelled (1) and (2) here:

$$\begin{aligned}
 E(\hat{\mathbf{y}} - \mathbf{y}) &= E(\mathbf{Z}\hat{\boldsymbol{\beta}} + \hat{M}(\mathbf{x1}, \mathbf{x2})) - \mathbf{Z}\boldsymbol{\beta} - M(\mathbf{x1}, \mathbf{x2}) \\
 &= E(\mathbf{Z}\hat{\boldsymbol{\beta}} + \mathbf{S}(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}})) - \mathbf{Z}\boldsymbol{\beta} - M(\mathbf{x1}, \mathbf{x2}) \\
 &= \mathbf{Z}E(\hat{\boldsymbol{\beta}}) + \mathbf{S}(\mathbf{Z}\boldsymbol{\beta} + M(\mathbf{x1}, \mathbf{x2}) - \mathbf{Z}E(\hat{\boldsymbol{\beta}})) - \mathbf{Z}\boldsymbol{\beta} - M(\mathbf{x1}, \mathbf{x2}) \\
 &= (\mathbf{I} - \mathbf{S})\mathbf{Z}\underbrace{(E(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta})}_1 + \underbrace{\mathbf{S}M(\mathbf{x1}, \mathbf{x2}) - M(\mathbf{x1}, \mathbf{x2})}_2. \quad (5.10)
 \end{aligned}$$

Asymptotic expressions for these component biases will be presented for each of the models in Sections 5.2.1-5.2.3 below.

### 5.2.1 Univariate SAM

Recall the results of Section 3.2.5 for the semiparametric model with a single smooth covariate and a single linear covariate. The difference between this model and the models under consideration here is that the number of linear terms is no longer restricted. Changing the notation of Section 3.2.5 slightly, the model becomes Equation 5.2. Similarly altering the notation of Equation 3.4 yields an expression for the asymptotic bias of  $\hat{\boldsymbol{\beta}}$  in this setting:

$$E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} | \mathbf{x1}, \mathbf{Z}) = -\frac{\mu_2(K)}{2}h^2 E(\text{var}(\mathbf{Z}_i | X1_i))^{-1} \text{cov}(\mathbf{Z}_i, m''(X1_i)) + o_p(h^2).$$

Since there is only one nonparametric term, the smoothing matrix associated with this term is the *centred* version of a univariate local linear smoothing matrix as described in Section 3.2.4. The use of a centred version reflects the inclusion of an ‘intercept’ term in the parametric component. This form is adopted to aid model comparison amongst the class of SAMs. The asymptotic bias of this component is presented in Section 3.2.4. Combining the two component biases yields the following expression for the asymptotic bias of a univariate-SAM fit using local linear regression:

$$E(\hat{Y}(X1_i, \mathbf{Z}_i) - Y(X1_i, \mathbf{Z}_i)) \approx \frac{h^2 \mu_2(K)}{2} \{m''(X1_i) - E(m''(X1_i)) - E(\text{var}(\mathbf{Z}_i|X1_i))^{-1} \text{cov}(\mathbf{Z}_i, m''(X1_i))\}.$$

It follows that the asymptotic bias of this model fit when the covariates are independent is given by:

$$E(\hat{Y}(X1_i, \mathbf{Z}_i) - Y(X1_i, \mathbf{Z}_i)) = \frac{h^2 \mu_2(K)}{2} (m''(X1_i) - E(m''(X1_i)))$$

Note especially the property that the asymptotic bias of the nonparametric component is as if the linear component was known. This property applies to any number of linear terms.

### 5.2.2 Additive SAM

In the case of the model  $E(\mathbf{y}) = m_1(\mathbf{x}1) + m_2(\mathbf{x}2) + \mathbf{Z}\beta$  the method of fitting proceeds as above (Equations 5.5 & 5.6) using the explicit definition of a bivariate additive model smoothing matrix defined in Section 3.2.6,  $\mathbf{S}_{AM}$  say. The two components of the additive model,  $m_1(\mathbf{x}1)$  and  $m_2(\mathbf{x}2)$ , can be estimated separately as:

$$\hat{m}_1(\mathbf{x}1) = \{\mathbf{I} - (\mathbf{I} - \mathbf{S}_1\mathbf{S}_2)^{-1}(\mathbf{I} - \mathbf{S}_1)\}(\mathbf{y} - \mathbf{Z}\hat{\beta})$$

$$\hat{m}_2(\mathbf{x}2) = \{\mathbf{I} - (\mathbf{I} - \mathbf{S}_2\mathbf{S}_1)^{-1}(\mathbf{I} - \mathbf{S}_2)\}(\mathbf{y} - \mathbf{Z}\hat{\beta}).$$

where  $\mathbf{S}_1$  &  $\mathbf{S}_2$  are the centred univariate smoothing matrices with respect to  $\mathbf{x}1$  and  $\mathbf{x}2$  respectively.

Opsomer & Ruppert (1999) is the only known treatment of the asymptotic properties of such a model fit and fortunately they use local linear regression. They point out that the approaches of Speckman (1988) and Carroll *et al.* (1997) do not generalise easily to the case where there is more than one nonparametric term entering additively. This supports their use (and indeed our own) of the backfitting paradigm to fit the model.

Opsomer & Ruppert (1999) only present the leading terms of the conditional bias and variance of the parameter estimates. In justifying the decision not to present explicit expressions of asymptotic bias they explain (pp. 7) ‘these expressions would be recursive as well as very complicated’. Although this is indeed true when there are more than two nonparametric additive terms, the case of two does lead to explicit expressions, as indeed the authors themselves imply in the previous paragraph where they acknowledge Hastie and Tibshirani (1990) as a source of explicit expressions for the fits.

Here we derive explicit definitions of the conditional asymptotic bias of the Additive-SAM, using arguments appearing in the proof of Opsomer and Ruppert’s (1999) expressions for the Univariate-SAM. Recall that Equation 5.10 showed that the overall bias of these semiparametric fits is the sum of biases of the parametric and nonparametric component individually.

Consider first the semiparametric component’s bias, i.e.

$$\begin{aligned}
 E(\hat{\beta} | \mathbf{Z}, \mathbf{x1}, \mathbf{x2}) - \beta &= (\mathbf{Z}^T (\mathbf{I} - \mathbf{S}_{AM}) \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{I} - \mathbf{S}_{AM}) E(\mathbf{y} | \mathbf{Z}, \mathbf{x1}, \mathbf{x2}) - \beta \\
 &= (\mathbf{Z}^T (\mathbf{I} - \mathbf{S}_{AM}) \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{I} - \mathbf{S}_{AM}) (M(\mathbf{x1}, \mathbf{x2}) + \mathbf{Z}\beta) - \beta \\
 &= (\mathbf{Z}^T (\mathbf{I} - \mathbf{S}_{AM}) \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{I} - \mathbf{S}_{AM}) (m_1(\mathbf{x1}) + m_2(\mathbf{x2})) \\
 &= \left( \frac{1}{n} \mathbf{Z}^T (\mathbf{I} - \mathbf{S}_{AM}) \mathbf{Z} \right)^{-1} \frac{1}{n} \mathbf{Z}^T (\mathbf{I} - \mathbf{S}_{AM}) (m_1(\mathbf{x1}) + m_2(\mathbf{x2})).
 \end{aligned}$$

Consider first,

$$\frac{1}{n} \mathbf{Z}^T (\mathbf{I} - \mathbf{S}_{AM}) \mathbf{Z} = \frac{1}{n} \mathbf{Z}^T \mathbf{Z} - \frac{1}{n} \mathbf{Z}^T \mathbf{S}_{AM} \mathbf{Z}.$$

By analogy with the proof of Theorem of 2.1 in Ruppert and Wand (1994),

$$\begin{aligned}
 \frac{1}{n} \mathbf{Z}^T \mathbf{S}_{AM} \mathbf{Z} &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \mathbf{Z}_i \mathbf{Z}_j^T [s_{X1_i, X2_i}]_j \\
 &\approx E(\mathbf{Z}_i E(\mathbf{Z}_i^T | X1_i, X2_i))
 \end{aligned} \tag{5.11}$$

where  $\mathbf{Z}_i$  is the  $i$ th row of  $\mathbf{Z}$  and  $[s_{X1_i, X2_i}]_j$  is the  $j$ th element of the  $i$ th row of

$\mathbf{S}_{AM}$ , the notation indicating that the smoothing matrix returns fitted values at the  $n$  observed design points. Hence,

$$\begin{aligned} \left( \frac{1}{n} \mathbf{Z}^T (\mathbf{I} - \mathbf{S}_{AM}) \mathbf{Z} \right)^{-1} &= (E(\mathbf{Z}_i(\mathbf{Z}_i - E(\mathbf{Z}_i|X1_i, X2_i))^T))^{-1} + o_p(h^2) \\ &= \text{var}(\mathbf{Z}_i|X1_i, X2_i)^{-1} + o_p(h^2). \end{aligned}$$

The second component,  $\frac{1}{n} \mathbf{Z}^T (\mathbf{I} - \mathbf{S}_{AM})(m_1(\mathbf{x1}) + m_2(\mathbf{x2}))$ , involves the bias of an additive model fit, described at length in Section 3.2.6. If the covariates  $X1$  and  $X2$  are independent then

$$\begin{aligned} (\mathbf{I} - \mathbf{S}_{AM})(m_1(\mathbf{x1}) + m_2(\mathbf{x2})) &\approx \frac{h_1^2}{2} \mu_2(K_1)(m_1''(\mathbf{x1}) - E(m_1''(\cdot))) + \\ &\quad \frac{h_2^2}{2} \mu_2(K_2)(m_2''(\mathbf{x2}) - E(m_2''(\cdot))). \end{aligned}$$

Hence,

$$\begin{aligned} \frac{1}{n} \mathbf{Z}^T (\mathbf{I} - \mathbf{S}_{AM})(m_1(\mathbf{x1}) + m_2(\mathbf{x2})) &\approx \frac{h_1^2}{2} \mu_2(K_1) E(\mathbf{Z}_i m_1''(X1_i) - \mathbf{Z}_i E(m_1''(\cdot))) + \\ &\quad \frac{h_2^2}{2} \mu_2(K_2) E(\mathbf{Z}_i m_2''(X2_i) - \mathbf{Z}_i E(m_2''(\cdot))) \end{aligned}$$

Noting that  $E(\mathbf{Z}_i m_1''(X1_i)) - E(\mathbf{Z}_i) E(m_1''(\cdot)) = \text{cov}(\mathbf{Z}_i, m_1''(X1_i))$  and combining the two components yields:

$$\begin{aligned} E(\hat{\beta} | \mathbf{Z}, \mathbf{x1}, \mathbf{x2}) - \beta &\approx \text{var}(\mathbf{Z}_i | X1_i, X2_i)^{-1} \left( \frac{h_1^2}{2} \mu_2(K_1) \text{cov}(\mathbf{Z}_i, m_1''(X1_i)) + \right. \\ &\quad \left. \frac{h_2^2}{2} \mu_2(K_2) \text{cov}(\mathbf{Z}_i, m_2''(X2_i)) \right) \end{aligned}$$

If we further assume that independence extends to *all* the covariates, i.e.  $Z1, \dots, Zp, X1, X2$ , then the bias of the parametric component disappears leaving only the bias associated with the nonparametric term (Equation 5.10), i.e.

$$E(\hat{\mathbf{y}} | \mathbf{Z}, \mathbf{x1}, \mathbf{x2}) = E(\hat{m}_1(\mathbf{x1}) + \hat{m}_2(\mathbf{x2}) | \mathbf{Z}, \mathbf{x1}, \mathbf{x2}) - m_1(\mathbf{x1}) - m_2(\mathbf{x2}).$$

which is the bias of an additive model fit, familiar from Section 3.2.6.

### 5.2.3 Bivariate SAM

Fitting the model  $E(\mathbf{y}) = m(\mathbf{x1}, \mathbf{x2}) + \mathbf{Z}\beta$  requires the smoothing matrix,  $\mathbf{S}_{BS}$  say, of a local linear bivariate smooth described in Section 3.2.7. The fitted values of the bivariate-SAM are then calculated using Equation 5.7.

Expressions for the asymptotic bias of this model fit are derived using the same strategy as in Section 5.2.2. For this model, clearly  $\mathbf{S}_{AM}$  is replaced by  $\mathbf{S}_{BS}$  and consequently  $[s_{X1_i, X2_i}]_j$  in Equation 5.11 is the  $j$ th element of the  $i$ th row of  $\mathbf{S}_{BS}$ . The properties of the local linear bivariate smoother, as described in Ruppert and Wand (1994) and summarised in Section 3.2.7, lead to

$$\begin{aligned} \left(\frac{1}{n}\mathbf{Z}^T(\mathbf{I} - \mathbf{S}_{BS})\mathbf{Z}\right)^{-1} &\approx (E(\mathbf{Z}_i(\mathbf{Z}_i - E(\mathbf{Z}_i|X1_i, X2_i))^T))^{-1} \\ &= \text{var}(\mathbf{Z}_i|X1_i, X2_i)^{-1} \end{aligned}$$

and

$$\begin{aligned} (\mathbf{I} - \mathbf{S}_{BS})(m(\mathbf{x1}, \mathbf{x2})) &\approx \frac{1}{2}\mu_2(K) \left( h_1^2 \left\{ \frac{\partial^2 m}{\partial X1^2} \Big|_{\mathbf{x1}, \mathbf{x2}} - E\left(\frac{\partial^2 m}{\partial X1^2}\right) \right\} + \right. \\ &\quad \left. h_2^2 \left\{ \frac{\partial^2 m}{\partial X2^2} \Big|_{\mathbf{x1}, \mathbf{x2}} - E\left(\frac{\partial^2 m}{\partial X2^2}\right) \right\} \right) \end{aligned}$$

Hence,

$$\begin{aligned} \frac{1}{n}\mathbf{Z}^T(\mathbf{I} - \mathbf{S}_{BS})m(\mathbf{x1}, \mathbf{x2}) &\approx \frac{1}{2}\mu_2(K) \left( h_1^2 E \left\{ \mathbf{Z}_i \frac{\partial^2 m}{\partial X1^2} \Big|_{\mathbf{x1}, \mathbf{x2}} - E\left(\frac{\partial^2 m}{\partial X1^2}\right) \right\} + \right. \\ &\quad \left. h_2^2 E \left\{ \mathbf{Z}_i \frac{\partial^2 m}{\partial X2^2} \Big|_{\mathbf{x1}, \mathbf{x2}} - E\left(\frac{\partial^2 m}{\partial X2^2}\right) \right\} \right) \\ &= \frac{1}{2}\mu_2(K) \left( h_1^2 \text{cov} \left( \mathbf{Z}_i, \frac{\partial^2 m}{\partial X1^2} \Big|_{X1_i, X2_i} \right) + \right. \end{aligned}$$

$$h_2^2 \text{cov} \left( \mathbf{Z}_i, \frac{\partial^2 m}{\partial X_2^2} \middle|_{X_{1i}, X_{2i}} \right)$$

Combining these results give an asymptotic approximation for the bias of the parameter estimate as:

$$E(\hat{\beta} | \mathbf{Z}, \mathbf{x}_1, \mathbf{x}_2) - \beta \approx \frac{1}{2} \mu_2(K) \text{var}(\mathbf{Z}_i | X_{1i}, X_{2i})^{-1} \left( h_1^2 \text{cov} \left( \mathbf{Z}_i, \frac{\partial^2 m}{\partial X_1^2} \middle|_{X_{1i}, X_{2i}} \right) + h_2^2 \text{cov} \left( \mathbf{Z}_i, \frac{\partial^2 m}{\partial X_2^2} \middle|_{X_{1i}, X_{2i}} \right) \right)$$

As before, independence amongst the covariates leads to the bias of the parametric component disappearing leaving only the bias associated with the nonparametric term (Equation 5.10), i.e.

$$E(\hat{y} | \mathbf{Z}, \mathbf{x}_1, \mathbf{x}_2) = E(\hat{m}(\mathbf{x}_1, \mathbf{x}_2) | \mathbf{Z}, \mathbf{x}_1, \mathbf{x}_2) - m(\mathbf{x}_1, \mathbf{x}_2).$$

which is the bias of a bivariate smooth model fit, familiar from Section 3.2.7.

### 5.2.4 Summary

This section has presented the asymptotic bias results for the SAM class of models. For each model, the bias comprised two components, the bias of the parameter estimates and the bias of the nonparametric component fit. In each case the former term involved the covariance of the linear covariates with the second derivative of the underlying regression function and the conditional variance of the linear covariates conditional on the nonparametric covariates observed values. Hence, dependency between the covariates impacts on the asymptotic bias of the parameter estimates and thus the entire model fit.

If, however, the covariates were independent, then in each case it was observed that the asymptotic bias of the model fit reduced to the bias associated with the nonparametric term only. The three forms of the nonparametric term were well documented in Chapter 3 and we therefore propose to use the attractive ‘cancelling’ properties of these biases to propose and investigate methods of inference

amongst this extended class when the covariates are independent in the next section.

### 5.3 Model Comparisons

The previous section has highlighted similarities between the class of models described in this chapter and the class considered extensively in Chapter 3, namely:

- ◊ explicit fits using smoothing matrices exist and are computationally feasible;
- ◊ under certain conditions, asymptotic bias results indicate the potential for biases to cancel.

These properties suggest that the methods of inference described in Chapter 3 may be appropriate for use in this setting. The extension of the methodology described in Section 3.5 is straightforward since all the information necessary to compare models is contained within the smoothing matrices which yield the model fits to be compared.

One difference between methods of inference for the bivariate class of Chapter 3 and the extended class here is that RSS based estimators of error variance must be used rather than difference based methods. Chapter 2 described extensions of difference based estimators of  $\sigma^2$  from the univariate setting to the bivariate setting. In the current setting though, the number of covariates is limitless and therefore to generalise these results for use here would require difference based techniques that accommodate an unlimited number of covariates. This is not a feasible approach since we have already seen how quickly concepts such as local neighbourhood and neighbouring points become complicated as the dimensions increase. Fortunately the advantage of RSS based estimators is that they do generalise easily. This decision is also supported by the results of the simulations in Chapter 4 which showed that, when used with an appropriate level of smoothing, RSS based estimators performed well in all settings.



Another difference between this setting and that of Chapter 3 also relates to the increased number of potential covariates. As more covariates enter the models, more data are required to make sensible estimates of the underlying regression function. Furthermore, since the methods of estimation and inference rely heavily on matrix manipulations the number of observations to be processed is a serious consideration. For instance, a sample size of 100 in a bivariate setting corresponds to a sample size of 1000 with three covariates, 10000 with four and so on. This is a consequence of the ‘curse of dimensionality’ and although SAMs partially counter this via their additive nature, more data is required the greater the number of parameters that are included in the model. This raises serious questions concerning the computational costs of performing model inference, especially when simulation studies are the prime vehicle by which methods are assessed.

Methods of inference via model comparisons based on RSSD and CFV comparisons of fitted values are explored in the next section. Three model comparisons are investigated, namely:

1. linear vs. univariate SAM
2. univariate SAM vs. additive SAM
3. additive SAM vs. bivariate SAM

In each case three covariates enter the model and the comparisons define a natural progression from the most restrictive to the most complex model. Extensions beyond three covariates involve extra linear terms and therefore will show similar behaviour to those considered here.

Simulated data over both regular grid and uniform-random designs are used to examine the tests’ performances. In the case of the regular grid designs the smoothing matrices (and associated terms) need only be calculated once for each model comparison and therefore a variety of settings and scenarios is considered. Since the random designs require substantially more computational resources the simulations focus on a selection of settings which highlight the properties of the different test approaches in this context.

## 5.4 Simulation Study

The regression functions  $M(X1, X2) + Z\beta$  corresponding to each of these models and used to simulate data, are:

<b>linear:</b>	$X1 + X2 + Z/2$
<b>univariate SAM:</b>	$1 + X1 - 0.75 \exp(-0.5(X1 - 0.5)^2/0.01) + Z/2$
<b>additive SAM:</b>	$2 + X1 + X2 - 0.75 \exp(-0.5(X1 - 0.5)^2/0.01) +$ $0.75 \exp(-0.5(X2 - 0.5)^2/0.01) + Z/2$
<b>bivariate SAM:</b>	$X2 - X1 + \Psi(X1, X2, 0.3, 0.3, 0.1, 0.1, 0) -$ $\Psi(X1, X2, 0.7, 0.7, 0.1, 0.1, 0) + Z/2$

where  $\Psi(x_1, x_2, \mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$  denotes the probability density function of a bivariate normal distribution. Each regression function is scaled such that the range of values over  $[0, 1]^2$  is 1 before random noise is added.

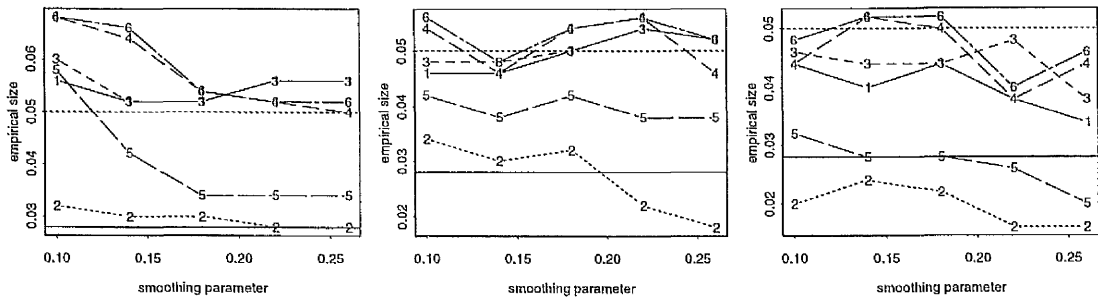
As with previous simulation studies, a variety of data conditions and test approaches will be considered. These are listed in Table 5.1. Note that in addition to considering the sample size and error variance effects, tests proceed using a range of smoothing parameters. In some instances the true value of  $\sigma^2$  is used in the test procedure. This examines the behaviour of the comparison of model fit statistics apart from the effect of estimating  $\sigma^2$ . The inclusion of RSS estimators of  $\sigma^2$ , using both the full model fit used in the model comparisons and an undersmoothed ( $h/2$ ) version, examine the realistic scenario of an unknown  $\sigma^2$ .

### 5.4.1 Regular grid designs

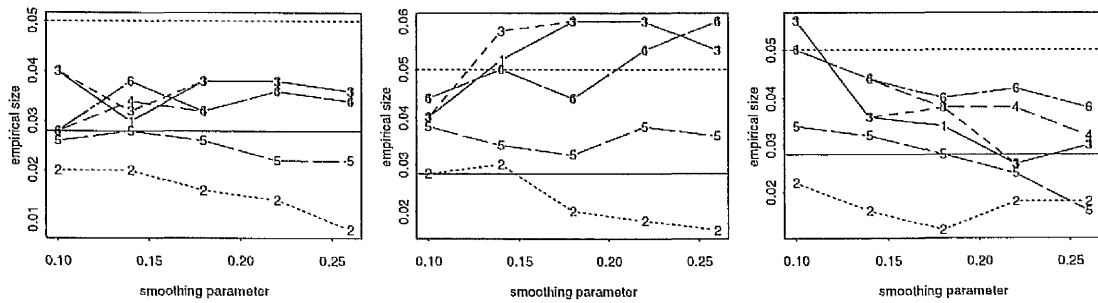
#### Size results

Empirical sizes at the  $\alpha = 0.05$  level for each test and setup are presented in Figures 5.1-5.2. The data were simulated from the regression function corresponding to the reduced model in each case. As previously, a 99% confidence interval based on the  $\text{Bi}(500, 0.05)$  distribution is shown to help assess how consistent the empirical sizes are with the specified significance level.

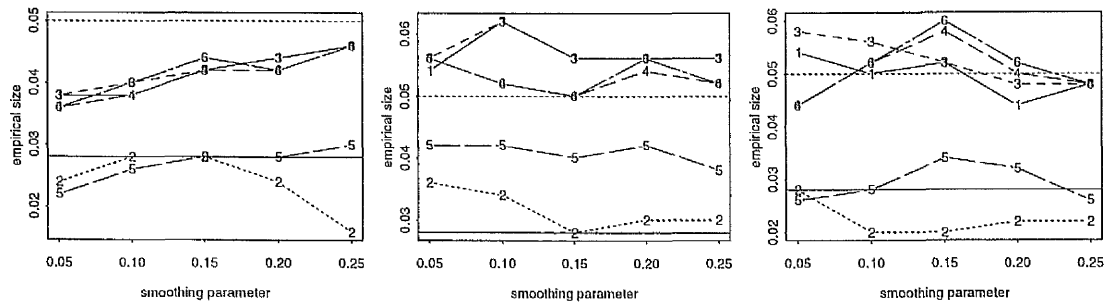
linear vs. uni.sp :  $n = 343$  Sigma = 0.1    uni.sp vs. add.sp :  $n = 343$  Sigma = 0.1    add.sp vs. bi.sp :  $n = 343$  Sigma = 0.1



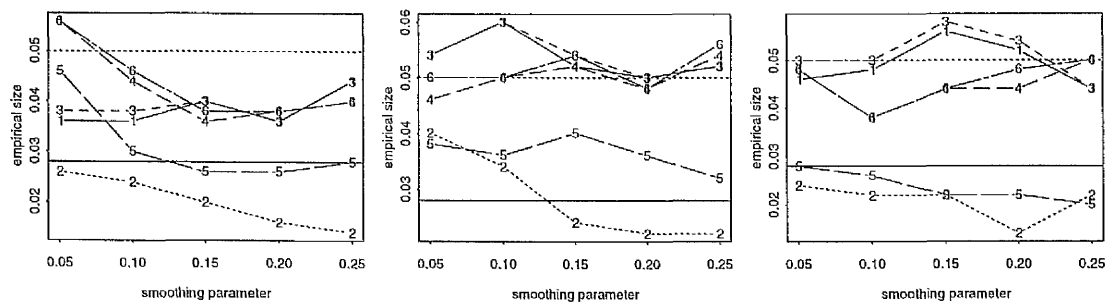
linear vs. uni.sp :  $n = 343$  Sigma = 0.2    uni.sp vs. add.sp :  $n = 343$  Sigma = 0.2    add.sp vs. bi.sp :  $n = 343$  Sigma = 0.2



linear vs. uni.sp :  $n = 512$  Sigma = 0.1    uni.sp vs. add.sp :  $n = 512$  Sigma = 0.1    add.sp vs. bi.sp :  $n = 512$  Sigma = 0.1

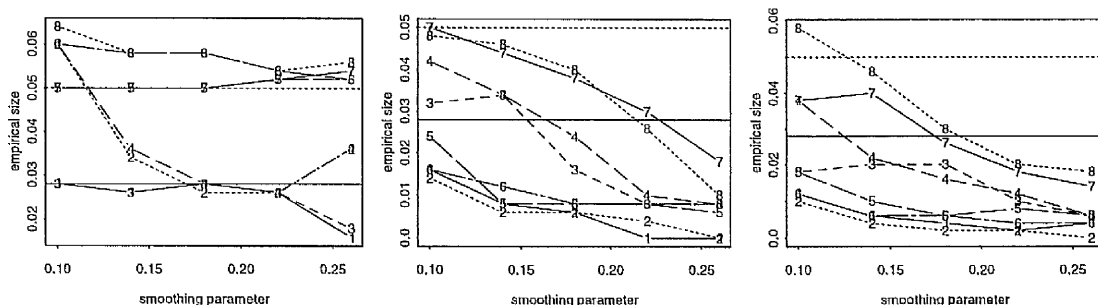


linear vs. uni.sp :  $n = 512$  Sigma = 0.2    uni.sp vs. add.sp :  $n = 512$  Sigma = 0.2    add.sp vs. bi.sp :  $n = 512$  Sigma = 0.2

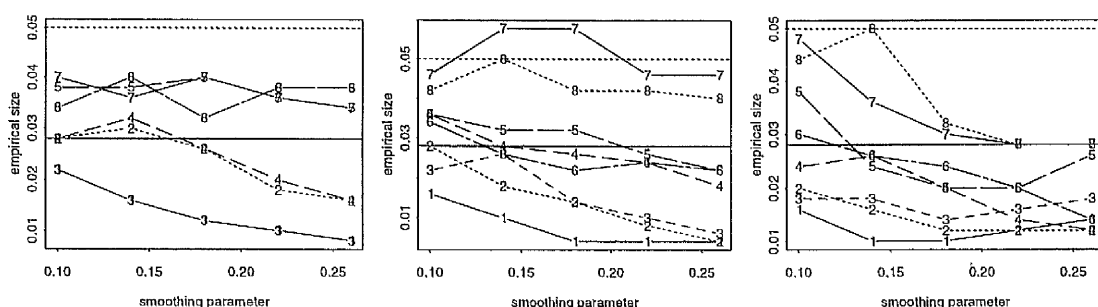


**Figure 5.1.** Empirical sizes, each based on 500 simulated data sets over regular grid designs. Model comparison tests use the known value of  $\sigma^2$ . Legend (Model comparison/Ref. Distn.): 1-CFV/QF; 2-CFV/ $\chi^2$ ; 3-CFV/ $\chi^2$ .corr; 4-RSSD/QF; 5-RSSD/ $\chi^2$ ; 6-RSSD/ $\chi^2$ .corr.

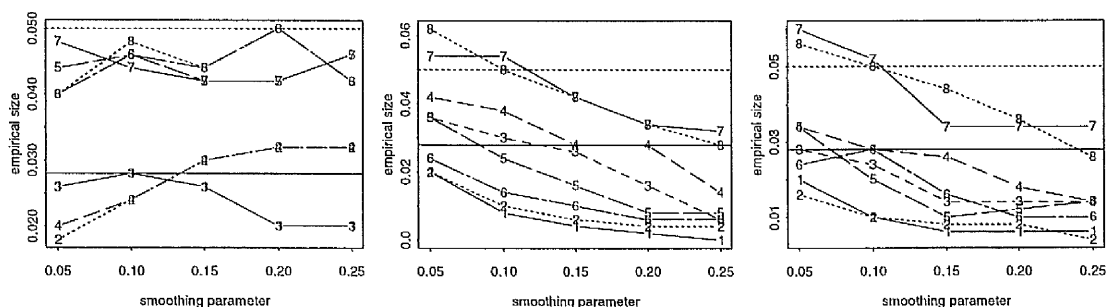
linear vs. uni.sp :  $n = 343$  Sigma = 0.1    uni.sp vs. add.sp :  $n = 343$  Sigma = 0.1    add.sp vs. bi.sp :  $n = 343$  Sigma = 0.1



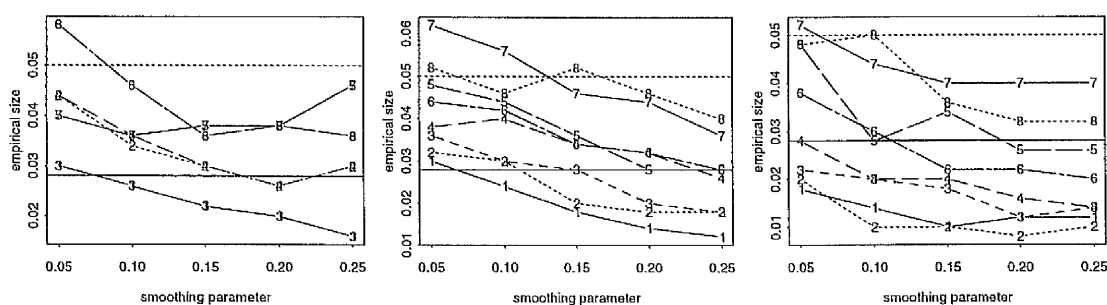
linear vs. uni.sp :  $n = 343$  Sigma = 0.2    uni.sp vs. add.sp :  $n = 343$  Sigma = 0.2    add.sp vs. bi.sp :  $n = 343$  Sigma = 0.2



linear vs. uni.sp :  $n = 512$  Sigma = 0.1    uni.sp vs. add.sp :  $n = 512$  Sigma = 0.1    add.sp vs. bi.sp :  $n = 512$  Sigma = 0.1



linear vs. uni.sp :  $n = 512$  Sigma = 0.2    uni.sp vs. add.sp :  $n = 512$  Sigma = 0.2    add.sp vs. bi.sp :  $n = 512$  Sigma = 0.2



**Figure 5.2.** Empirical sizes, each based on 500 simulated data sets over regular grid designs. Legend for test approaches (Model comparison/Ref. Distn./ $\hat{\sigma}^2$ ): 1-CFV/F/RSS; 2-RSSD/F/RSS; 3-CFV/F/RSS(us); 4-RSSD/F/RSS(us); 5-CFV/F.cor/RSS; 6-RSSD/F.cor/RSS; 7-CFV/F.cor/RSS(us); 8-RSSD/F.cor/RSS(us).

**Table 5.1.** Settings and test approaches investigated via a simulation study the results of which are shown in Figures 5.1 & 5.2

Data conditions	
design space	regular grid over $[0, 1]^3$
design space	uniform-random over $[0, 1]^3$
sample size	343 ( $7^3$ ) design points
	512 ( $8^3$ ) design points (grid only)
error distribution	Normal (zero mean)
error variance	0.01
	0.04
number of simulations	500 - regular
	250 - random
Test approaches	
model comparisons	Linear vs. Uni. SAM
	Uni. SAM vs. Add. SAM
	Add. SAM vs. Biv. SAM
smoothing parameter	0.1, 0.14, 0.18, 0.22, 0.26 (n=343)
	0.05, 0.1, 0.15, 0.2, 0.25 (n=512)
variance estimator	NONE (true value used)
	RSS based (incl. undersmoothing)
comparisons of model fits	RSSD and CFV
reference distribution	$F$ , corrected $F$

Figure 5.1 shows results when the true value of  $\sigma^2$  is used in the test procedures. In most cases both the RSSD and the CFV tests return consistent results. The only exception is when an (uncorrected)  $\chi^2$  distribution is used to calculate the p-value, in which case the sizes are too low, i.e. conservative tests.

Figure 5.2 shows the results when  $\sigma^2$  is estimated from the full model fit. From these results there is clear evidence that undersmoothing the full model fit to estimate  $\sigma^2$  and using a corrected  $F$  reference distribution is the best way to proceed, since this is the approach which returns the most consistent results. Problems do arise, however, with this approach when too big a smoothing parameter is used (especially at the smaller sample size) since the bias related deterioration in the methods is proportional to the smoothing parameter used. Despite this inevitable deterioration, the approach is consistent over a range of smoothing parameters,

indeed nearly all of the values considered here with sample size of 512.

The next best approach of undersmoothed  $\hat{\sigma}^2$  with an uncorrected F reference distribution reveals that the estimation of  $\sigma^2$  is the crucial factor in this setting. Other comparisons reveal that the CFV approach is more able to minimise the effect of the increasing bias at higher values of  $h$  than the RSSD comparison which is consistent with the asymptotic results for these fits.

### Power results

Power results for the CFV & RSSD comparisons with an undersmoothed RSS based estimator of  $\sigma^2$  using a corrected F reference distribution are shown in Table 5.2. These approaches were investigated for power since they returned the most consistent size results. In each model comparison the same regression functions as those used above were employed but the level of error variance was adjusted to return informative powers, i.e. away from the two extremes of zero and one.

The power results show that there is a tendency for the RSSD comparisons to have slightly higher power than the equivalent CFV comparisons with a maximum difference of 15%. The trend as the smoothing parameter increases is that the power of tests decrease with the exception of the 'linear vs. univariate SAM' comparison. This reflects the lack of bias in the fits when the regression function is linear.

### 5.4.2 Random designs

Given the computational effort required to simulate test results over random designs, 250 data sets of sample size  $n=343$  are considered here. The other settings, as listed in 5.1, are used here.

#### Size results

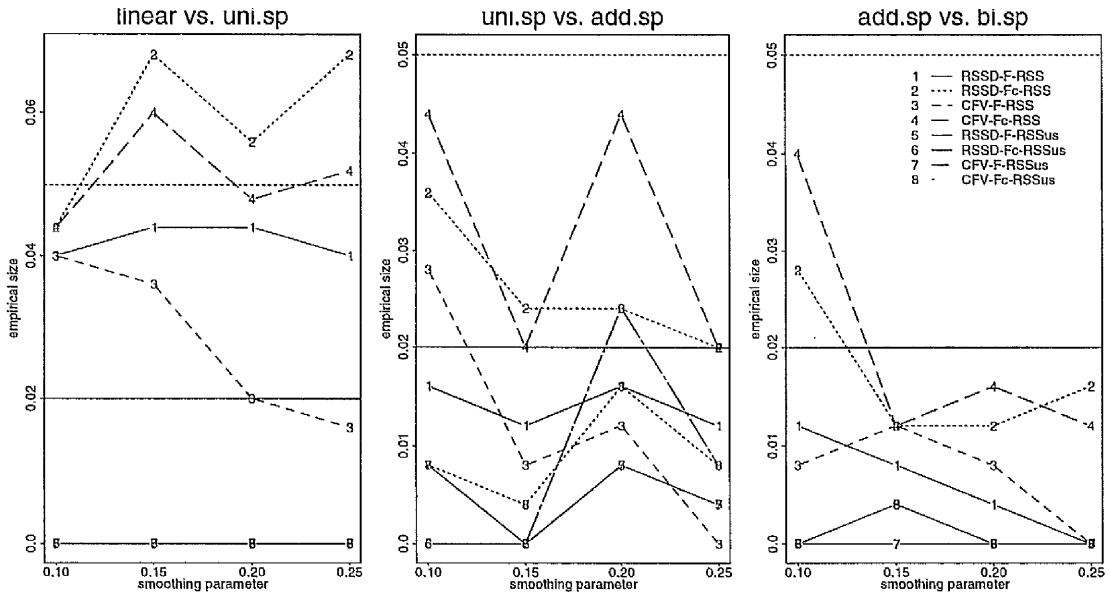
Figure 5.3 shows the results of 250 simulations which explore the size of the tests over random designs. The plots show that the best approaches use either CFV or RSSD comparisons with an RSS based estimator of  $\sigma^2$  (not undersmoothed)

**Table 5.2.** Power results for the CFV & RSSD comparisons with an under-smoothed RSS based estimator of  $\sigma^2$  using a corrected F reference distribution. The data settings are as shown in Table 5.1 with the exception of the error variance which are stated here.

Red. Model	Full Model	$\sigma$	h	n=343		n=512	
				CFV	RSS	CFV	RSS
Linear	Uni. SAM	1.0	0.10	0.410	0.448	0.560	0.594
Linear	Uni. SAM	1.0	0.14	0.346	0.432	0.520	0.600
Linear	Uni. SAM	1.0	0.18	0.340	0.418	0.514	0.600
Linear	Uni. SAM	1.0	0.22	0.366	0.398	0.548	0.578
Linear	Uni. SAM	1.0	0.26	0.398	0.376	0.576	0.568
Uni. SAM	Add. SAM	0.1	0.10	0.746	0.816	0.890	0.920
Uni. SAM	Add. SAM	0.1	0.14	0.640	0.794	0.832	0.910
Uni. SAM	Add. SAM	0.1	0.18	0.594	0.708	0.788	0.880
Uni. SAM	Add. SAM	0.1	0.22	0.584	0.618	0.794	0.844
Uni. SAM	Add. SAM	0.1	0.26	0.584	0.564	0.806	0.788
Add. SAM	Biv. SAM	0.2	0.10	0.498	0.580	0.680	0.784
Add. SAM	Biv. SAM	0.2	0.14	0.362	0.500	0.558	0.704
Add. SAM	Biv. SAM	0.2	0.18	0.276	0.378	0.434	0.588
Add. SAM	Biv. SAM	0.2	0.22	0.198	0.308	0.336	0.476
Add. SAM	Biv. SAM	0.2	0.26	0.170	0.260	0.246	0.424

and referenced to a corrected F distribution. The trends are reminiscent of comparable plots in Figure 4.17, which investigated model comparisons amongst the bivariate class over random designs. In the case of a linear regression function, the performance is consistently good over the range of smoothing parameters for both the corrected and uncorrected F distribution. The Uni. SAM versus Add. SAM comparison is consistent throughout the range of smoothing parameters when the corrected distribution is employed. The comparison of an Add. SAM fit versus a Biv. SAM fit proves also to be the most challenging, returning consistent results only at the smallest smoothing parameter values.

Of particular interest, however, is the effect of undersmoothing the full model fit to obtain  $\hat{\sigma}^2$ . In all of the previous results it has been noted that the effect of undersmoothing is to reduce the bias in  $\hat{\sigma}^2$  which therefore inflates the value of the model comparison test statistic which in turn increases its empirical sizes in



**Figure 5.3.** Empirical sizes, each based on 250 simulated data sets over uniform-random designs.

the simulations. In this instance, however, the effect is in the opposite direction, i.e. the use of an ‘undersmoothed’ estimate of  $\sigma^2$  *decreases* the empirical size of the tests.

To understand this behaviour, recall the expectation of the RSS estimator of  $\sigma^2$  based on the full model fit, derived from Equation 2.3:

$$E\left(\frac{RSS}{df_{error}}\right) = \sigma^2 + \frac{\mathbf{b}_F^T \mathbf{b}_F}{n - \text{tr}\{2\mathbf{S}_F - \mathbf{S}_F^T \mathbf{S}_F\}}.$$

From this expression it is easy to see why it is of potential benefit to reduce the bias of the fitted values of the full model. However, reducing the  $\mathbf{b}_F^T \mathbf{b}_F$  term by decreasing the smoothing parameter also impacts on the denominator  $n - \text{tr}\{2\mathbf{S}_F - \mathbf{S}_F^T \mathbf{S}_F\}$ . To understand this recall that the approximate error degrees of freedom is an analogue of  $n$  minus the number of parameters estimated in a parametric model, i.e. it is a measure of model complexity. Since reducing the smoothing parameter allows increased flexibility in the fitted model its effect is to reduce the error degrees of freedom available to estimate  $\sigma^2$ .

The final impact of undersmoothing on the accuracy of the RSS estimator



of  $\sigma^2$  is therefore dependent on its effect on both numerator and denominator of  $\mathbf{b}_F^T \mathbf{b}_F / (n - \text{tr}\{2\mathbf{S}_F - \mathbf{S}_F^T \mathbf{S}_F\})$ . Providing the relative reduction in  $\mathbf{b}_F^T \mathbf{b}_F$  is *greater than* the relative reduction in the approximate degrees of freedom, then the ultimate effect is to increase the size of the test. If, however, the converse is true then the result would be as witnessed in Figure 5.3, a reduction in the size of the tests.

The reason this has occurred in this setting and not before is because of the sample size and distribution of the design points. The design space is defined by generating 343 random numbers from a  $U(0, 1)$  distribution along each of the three axes and combining them to define the coordinates of the observed design point. Hence the marginal distributions, 343 points between zero and one, are very dense and therefore the effect of altering the smoothing parameter, in this case halving it, has a more pronounced effect than if there were  $n = 100$  points (i.e. the largest random design sample size considered previously).

In the case of the regular designs of Section 5.4.1 even though a comparable, and greater, number of design points were considered, the marginal distributions were equivalent to those considered previously, i.e.  $7 \times 7 \times 7$  and  $8 \times 8 \times 8$  grids respectively. There a reduction in smoothing parameter dominated the bias rather than the error degrees of freedom as it did in Chapter 4.

### Power results

To complete this investigation of the tests' performances with data generated from random designs, a small simulation study investigating the power of certain tests was performed. Because of its favourable 'size' performance, an approach using  $h=0.15$  and with no undersmoothing to obtain the RSS based estimator of  $\sigma^2$  is used with reference to a corrected  $F$  distribution. The simulated data corresponds to the full model in each case and the regression function and error variance used were the same as in the regular grid power simulations of Section 5.4.1.

Table 5.3 lists the power results for the three model comparisons and two methods of comparing fitted values. The absolute magnitude of these results is a reflection of the functions and error variances used to generate the true (full) models, and therefore is not of intrinsic interest. Of interest is a comparison of

Red. Model	Full Model	Model Comparison	
		CFV	RSSD
Linear	Uni. SAM	0.060	0.076
Uni. SAM	Add. SAM	0.252	0.264
Add. SAM	Biv. SAM	0.264	0.320

Table 5.3. Power results: random designs.

the two approaches to model comparisons. As seen previously, there is little to distinguish the two, with perhaps some evidence of slightly higher power using the RSSD approach which corresponds to its conservative size properties highlighted above.

## 5.5 Discussion

This chapter has demonstrated how the methods of inference in Chapters 3 and 4 can be used amongst a much more general class of models. Restrictions remain with regard to the number of nonparametric terms, but the unlimited number of linear terms makes the class of models widely applicable.

Several consequences associated with the inclusion of an unlimited number of linear terms were noted. Firstly, bias properties of the model fits had to be derived. This required adapting recent work in the local linear semiparametric model context to the class of interest. This highlighted the effect of correlation amongst the covariates which enters the asymptotic bias expressions, an inevitable consequence of the additive form of the models. Computational consequences were also noted, since the increasing dimension of the models requires more and more data to yield reliable fits, which in turn increases the computational effort required to fit the models. Techniques involving ‘binning’ the data have been suggested in the context of model fitting (Härdle and Scott, 1992) and therefore a natural extension of this work is to incorporate binning into the methods of inference. Bowman and Azzalini (1997) are currently supplementing their ‘sm.regression’ library of S-Plus functions to permit binning in the comparisons involving univariate smooths. This suggests that the methods for multivariate

model inference developed here could also be modified using binning to cater for large number of observations.

The behaviour of the different approaches to model inference was reminiscent of the equivalent comparisons in Chapter 4. As such there is evidence to suggest that, in the case of a regular grid design, the combination of an undersmoothed RSS estimator of  $\sigma^2$  with a two moment corrected F distribution and a CFV model comparison is the best way to proceed. In the case of random designs, once again there is evidence that undersmoothing to obtain an estimate of  $\sigma^2$  is not preferable.

This has raised yet another reason to question the use of undersmoothing to estimate  $\sigma^2$  in the context of model comparisons when the designs are random. In Chapter 4 it was noted that undersmoothing increased the empirical sizes of the tests. In this chapter, however, it was observed that undersmoothing led to *decreases* in empirical sizes, i.e. *away* from the specified significance level.

This highlights the connection between the sample size and the properties of estimates derived from fits to the data. Since undersmoothing's effect on  $\hat{\sigma}^2$  is a result of its relative effect on both the bias and the error degrees of freedom of the model fit, of which the former is never known in practice, it is advisable not to employ undersmoothing when the design is of a random nature. Fortunately, the results also show that not undersmoothing yields consistent sizes when (the same) small smoothing parameter is used to compare the model fits and estimate  $\sigma^2$  using the residuals of the full model fit. This would appear to be the best strategy for minimising the effect of bias, whilst retaining a test statistic whose distribution can be well approximated.

Lastly, no attempt has been made to quantify the effect of introducing correlation amongst the covariates. A detailed study of these effects would require considerable computational resources, beyond the reach of this present work. For this serious consideration, guidance can be taken from the results of Section 4.3.3 with regard to the effect of correlation amongst the nonparametric terms. The asymptotic results show, however, that SAMs should be avoided when there is considerable dependency between linear and nonparametric covariates as this introduces bias into the parameter estimates. Note, however, that to achieve stable fits of models with an *additive* form requires a design space which has a

sufficient spread of observations and therefore severe dependency amongst the covariates inhibits estimation as well as inference with nonparametric models.

## Chapter 6

# Assessment of Eutrophication in the Firth of Clyde

### 6.1 Introduction & Background

This chapter describes an analysis of environmental data describing estuarine water quality in the west of Scotland. The analysis forms part of a collaborative project with Dr. Brian Miller of the Scottish Environment Protection Agency (SEPA) West Region. The research findings of particular interest to marine scientists are presented in Bock *et al.* (1999).

Although the overall quality of most of the estuarine and coastal waters in Scotland is good or excellent, areas still exist where water quality is unsatisfactory, mainly in estuaries and inshore coastal waters. This is due to the effects of discharges of domestic and industrial wastes, and historically polluted sediments. One important indicator of the quality of the marine environment is the potential and occurrence of eutrophication, defined as:

the enrichment of water by nutrients, especially compounds of nitrogen and/or phosphorus, causing an accelerated growth of algae and higher forms of plant life to produce an undesirable disturbance to the balance of organisms present in the water and to the quality of the water concerned (Urban Waste Water Treatment Directive (UWWTD) 1991).

Nutrients are mineral substances required for the nutrition of plants (specifically phytoplankton in marine waters); hypernitrification is the build-up of nutrients.

This chapter investigates the potential for eutrophication in the coastal waters of the Firth of Clyde, Scotland. Data collected over a 14 year period at a single sampling station (coded 'CMT7' and shown in Figure 6.1) are analysed. The aim of the analysis is to investigate changes in the trophic status and the potential for eutrophication at this site over this period.

Section 6.2 describes the data, including the sampling scheme and the primary variables of interest. Section 6.3 describes different analyses together with results which investigate a variety of aspects of the system. Throughout these analyses, nonparametric techniques are employed since many of the assumptions of parametric models do not hold for these data. Section 6.4 discusses several aspects of the analyses and Section 6.5 ends with a summary of the main conclusions drawn from the results.

## 6.2 The Data

### 6.2.1 The variables

The CMT7 sampling station is one of 40 profile (multiple-depth) sampling stations and 79 surface waters sampling stations in the Firth of Clyde. These sites are monitored to assess the impact of discharges of sewage and industrial wastes on marine water quality. As can be seen in Figure 6.1, the CMT7 station is in an area of dynamic water movement where estuarine waters mix with more saline waters from the Firth of Clyde. These conditions promote good plankton growth so the patterns observed at the station are prone to the consequences of eutrophication. Since, primarily, interest lies in understanding changes across time and depth, it is sufficient to focus attention on this single key site. Not only is its geography key but it also has the most data of any single site.

At the CMT7 station samples are collected from 1, 4, 7, 10, 20, 30, 50, 70 and 90 metre depths. Table 6.1 lists the variables measured on the water samples collected at each depth, the information noted at the time of sampling, and the variables calculated from those measured. Although each of these variables was

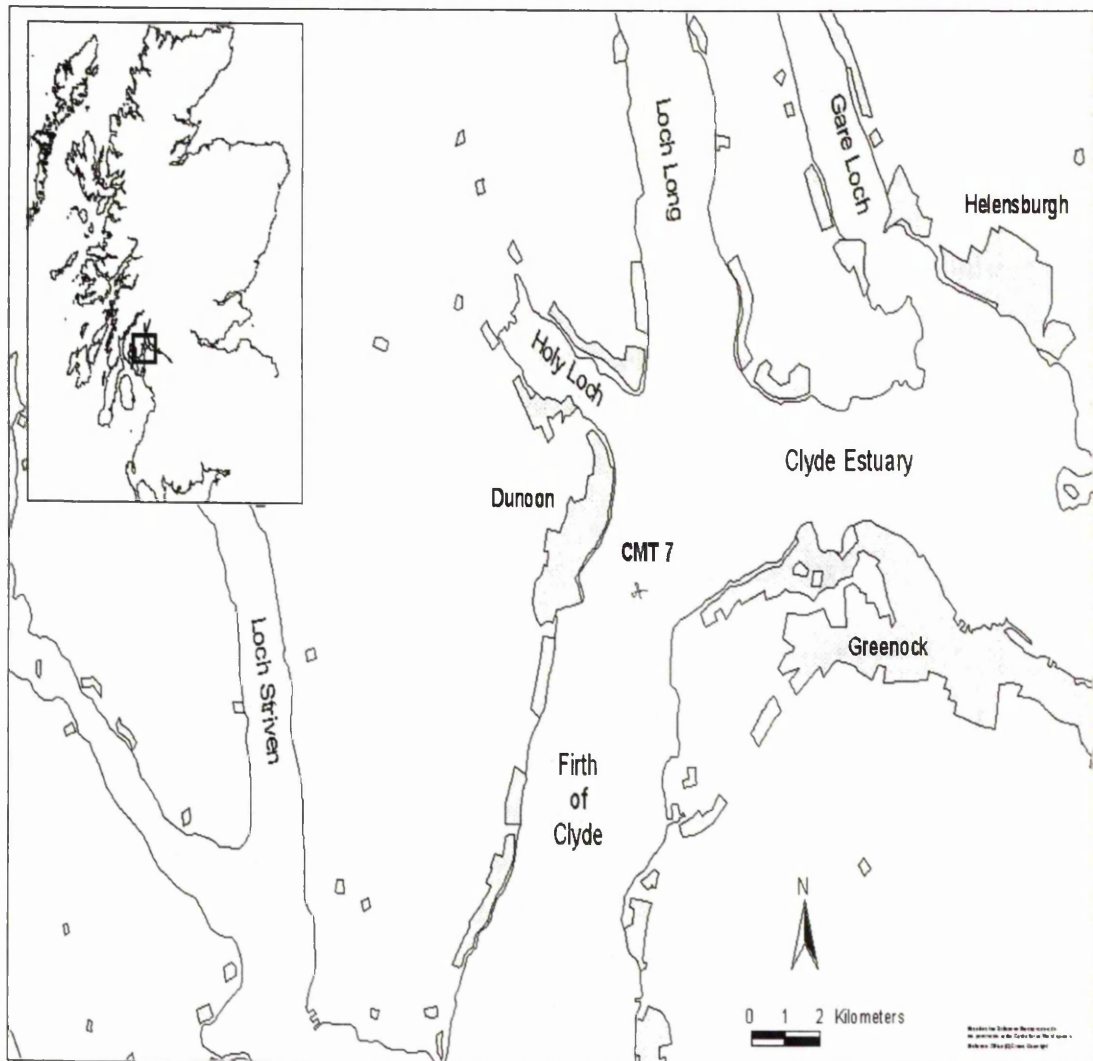


Figure 6.1. Map showing the location of the CMT7 sampling station

included in the analysis, this chapter will focus on the relationships between the nutrient compounds, *nitrate* and *phosphate*, and the measures of plankton levels, *chlorophyll a* and % *Dissolved Oxygen* (% D.O.).

**Table 6.1.** List of variables available for each water sample

Information Noted	Variables Measured	Variables Calculated
Date	Temperature ( $^{\circ}C$ )	Salinity (psu)
Time of Sampling	Conductivity (ratio)	Total Dissolved
Water Depth	Dissolved Oxygen ( $mg/l$ )	Nitrogen ( $mmol/m^3$ )
Time of Low Water	Nitrate ( $mmol/m^3$ )	% Dissolved Oxygen
Time of High Water	Nitrite ( $mmol/m^3$ )	
Tidal Range	Ammonia ( $mmol/m^3$ )	
	Phosphate ( $mmol/m^3$ )	
	Silicate ( $mmol/m^3$ )	
	Chlorophyll a ( $mg/m^3$ )	

## 6.2.2 Sampling frequency

SEPA West Region carries out routine water quality surveys in the Clyde Estuary, in the Firth of Clyde and in the adjacent sea lochs. The frequency of sampling at the CMT7 station during the period from March 1982 to April 1996 is shown in Table 6.2. The combination of bad weather and other operational constraints has led to an irregular sampling pattern, with the number of surveys ranging from 5 in the months of December to 22 in the Aprils. There is also an absence of data for much of 1985 when the survey vessel was temporarily out of service. The irregular distribution of sampling dates has added complexity to the data, so that standard time series methods cannot be used to investigate changes across time. For this reason exploratory data analysis using nonparametric techniques will be applied.



**Table 6.2.** Frequency of sampling at the ‘CMT7’ station

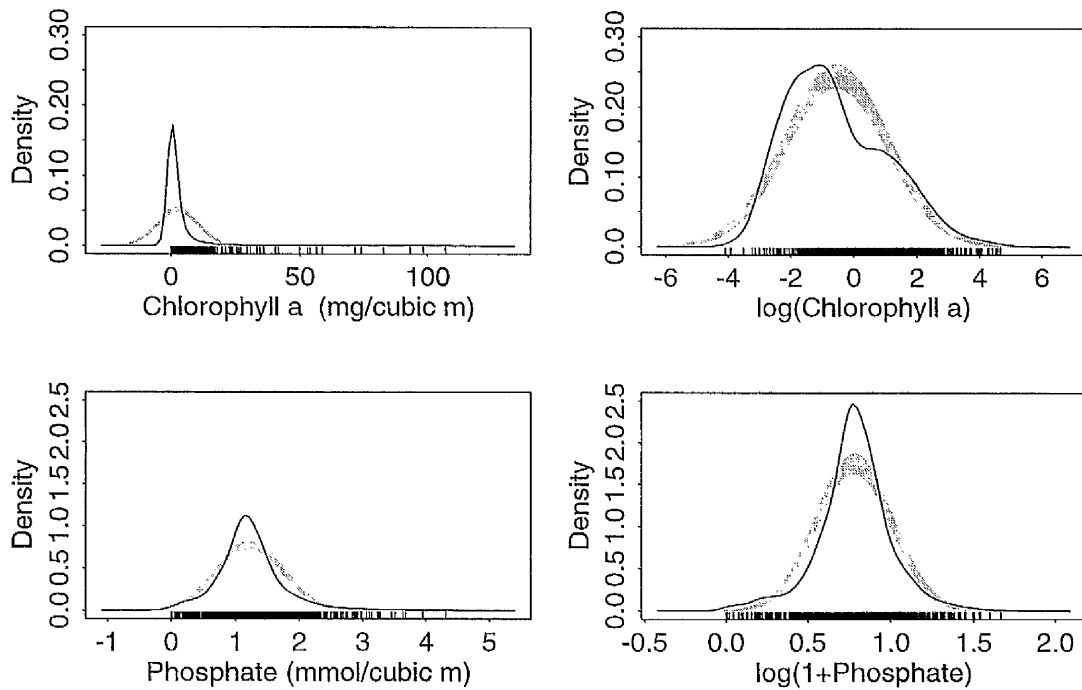
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
1982			***	**	*					*	*	**	10
1983	*			***	**		*	***	*	*	**		14
1984		*	**	*	**		*	*	*	***	*		13
1985									**	**	***		7
1986		**				*	**	*	**		*		9
1987	*	**		*			**	*	**		**		11
1988	***	*		**		**	*		*	*	*		12
1989	**	*		*	*		*	*	*	**	*	**	13
1990			**	**	*		**			***	*		11
1991	**	*	*	**	**	*	**		***		*		15
1992	*	**	*	*	**	*	*	*	*	**	*		14
1993		*	**	*	**	*	***	*	*	**	*		15
1994				*	*	*	**	**			***		10
1995		**	***	***	*	*	*	**	***	*	**	*	20
1996	*	***		**									6
Total	11	16	14	22	15	8	19	13	18	18	21	5	180

## 6.3 Analyses & Results

The analyses and results are presented in five subsections, each investigating a different aspect of the data. Sections 6.3.1-6.3.5, respectively, consider: the distributions of measured variables; changes in the concentrations of variables with depth; evidence for the presence of long term trends; seasonal patterns; and important aspects of the dynamics of the water system.

### 6.3.1 The distributions of measured variables

As a first step, the distributions of each of the variables listed in Table 6.1 were considered individually. As an illustration, the first column of Figure 6.2 shows estimates of the density function of the chlorophyll *a* and phosphate variables (tick marks indicate individual values upon which the estimates were based). These density estimates use Gaussian kernels and a smoothing parameter which minimise the Mean Integrated Squared Error (a measure of the difference between the estimated and true density) under the assumption that the underlying density is a normal distribution. Superimposed on each plot is a reference band indicating the likely position of density when the data are normal (Bowman and Azzalini, 1997 pp. 41). If the data are not normal, as is almost certainly the case here, the smoothing parameter will induce oversmoothing. Therefore, departures from



**Figure 6.2.** Density estimates of selected variables both before and after log transformations. Also shown are normal reference bands and tick marks indicating individual values.

the reference bands in Figure 6.2 are conservative in that they will tend to mask more marked departures of the 'true' density. The following departures from normality are therefore highlighted in Figure 6.2: the chlorophyll a values exhibit extreme positive skewness and the phosphate values less so. Positive skewness was observed in a high proportion of variables, which is common for environmental variables, which cannot take on negative values.

Logarithmic transformations were used to provide a more manageable scale for the values of the variables. The second column of Figure 6.2 displays the density estimates and normal reference bands for transformed chlorophyll a and phosphate values. Although there is still evidence of departures from normality, particularly in the body of the distributions, this is less than on the original scale, significantly so for chlorophyll a, and in both cases the tails are far more consistent with a normal distribution. Since no techniques requiring strictly

normal variables are used, the logarithmic transformations succeed in re-scaling the data for the purposes of these analyses. Similar transformations were made on the nitrogen compounds and silicate.

### 6.3.2 Changes in the concentrations of variables with depth

Figure 6.3 shows boxplots summarising the distributions of selected transformed variables at each of the nine depths separately. Several features are worth noting: firstly the range of concentrations observed for each variable is greatest in surface waters, with the ranges decreasing with depth; and secondly there is a relatively small change (relative to the spread of the values) in the median concentrations of the variables, especially across depths of 20-90m.

The observation that the largest spread in concentrations occurs in the surface waters is expected given the major influence of freshwater inputs on the composition of the surface layers in the water column. Other factors which contribute to this greater variability are the dynamic nature of the hydrographic and ecological processes taking place, in terms of inputs of organic and inorganic substances, vertical and horizontal migration of water bodies, uptake of nutrients by plankton, removal of dissolved species on particulates, etc. Much of the variation in surface waters is therefore attributable to seasonal effects (as demonstrated in Section 6.3.4).

The changes in the median concentrations of some variables, especially over the depth range from 20-90m, were smaller than expected. This has two main implications. Firstly, it suggests that overall the water column at depths greater than 20m is vertically well mixed. This suggests that samples at one depth, say 30m, may be taken as representative of the whole water column from 20 to 90m depth. Care must be taken, however, to discern whether this property is seasonally dependent. This is explored in Section 6.3.4.

Secondly, this information enables a data-based approach to be taken to the specification of a lower boundary for the depth at which a moored or towed constant monitoring device should be deployed. Figure 6.3 suggests that the most dynamic changes in water structure occur in the top 20m and therefore attention is best focused on observing the behaviour in this layer. This information is of

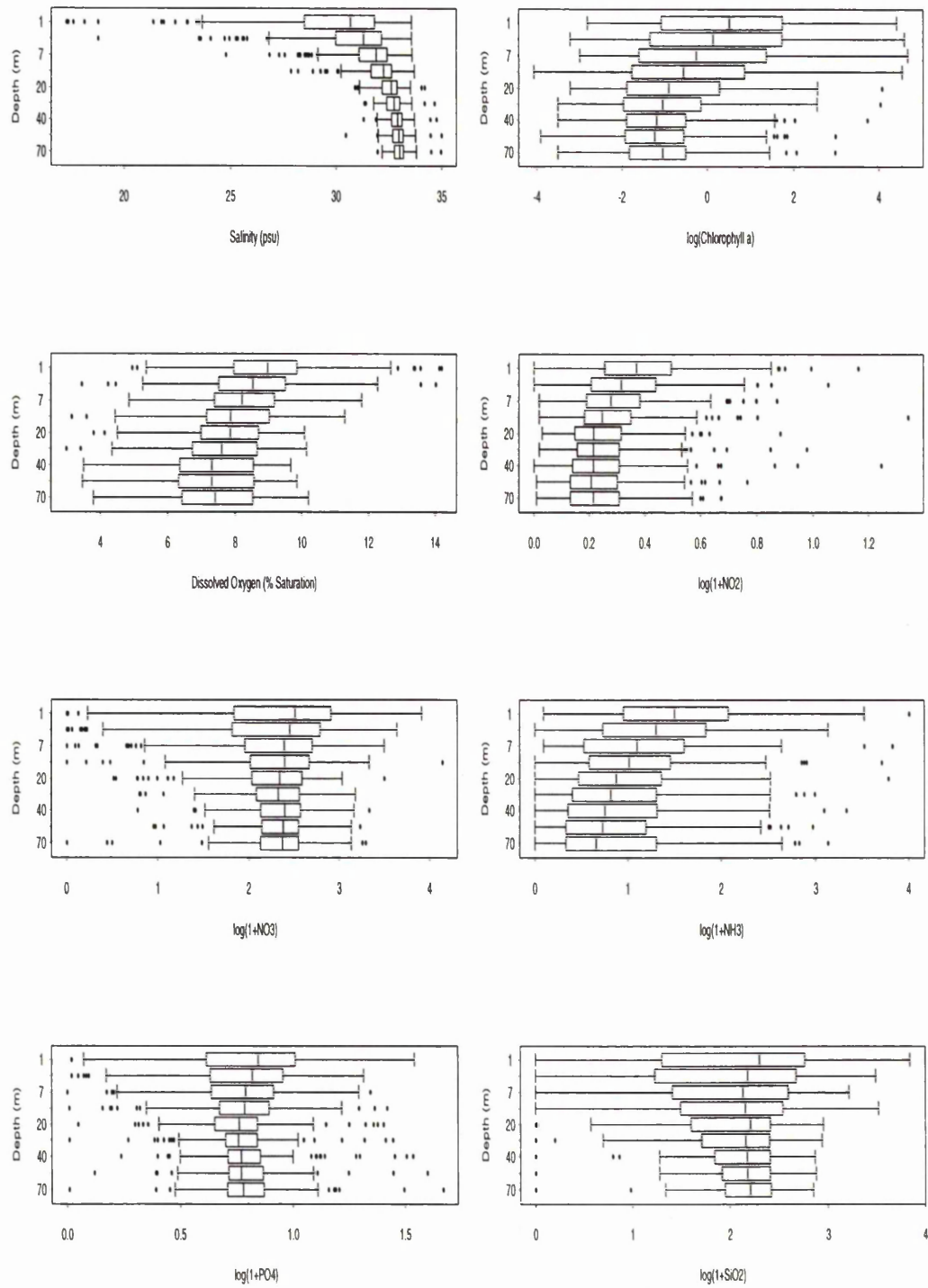


Figure 6.3. Distributions of selected variables with depth.

particular interest to scientists at SEPA West Region who have recently deployed a towed multi-variable monitoring device to measure coastal water quality.

### 6.3.3 Evidence for the presence of long term trends

When examining the changes in the variables' concentrations across time, care must be taken to differentiate between short term 'seasonal' variation (which is known to be present) and long term 'trends' in the values which may or may not be present. This section focuses on long term trend, Section 6.3.4 considers seasonal effects.

It was noted earlier that the irregular nature of the time series excludes the use of formal time series analyses. For this reason use will be made of nonparametric models to perform a descriptive analysis of the data. Although these models don't take into account the potentially correlated error structure in the data, they can be used to smooth out the 'noise' and therefore give some indication of underlying trends.

In the case of long term trends, it is reasonable to employ a smoothing technique which combines information across several years to estimate the mean level at a particular observation. By using a bandwidth of, say, 180 days with a normal kernel, data from approximately 1 year either side of each point is used to derive the local-linear kernel regression estimate. Hence, the seasonal effect is minimised by averaging over approximately two years of observations.

Figure 6.4 shows observations recorded during the 14-year period for dissolved oxygen and transformed nitrate and chlorophyll *a* concentrations in waters collected at 4, 20 and 70m depths. A local linear regression smoother has been applied to these observations using a bandwidth of 180 days. Superimposed on each plot is a reference band under the assumption of a constant trend. The reference band is centred around the mean value of the variable over the period and indicates the region where the nonparametric fit would likely lie if the underlying function was constant (see Bowman and Azzalini, 1997, pp.90).

Three features are worth noting. Firstly, the spread in the values of each variable decreases with depth, thus reflecting the features of Figure 6.3. Secondly there is no clear evidence of a long term trend with time for any variable. This

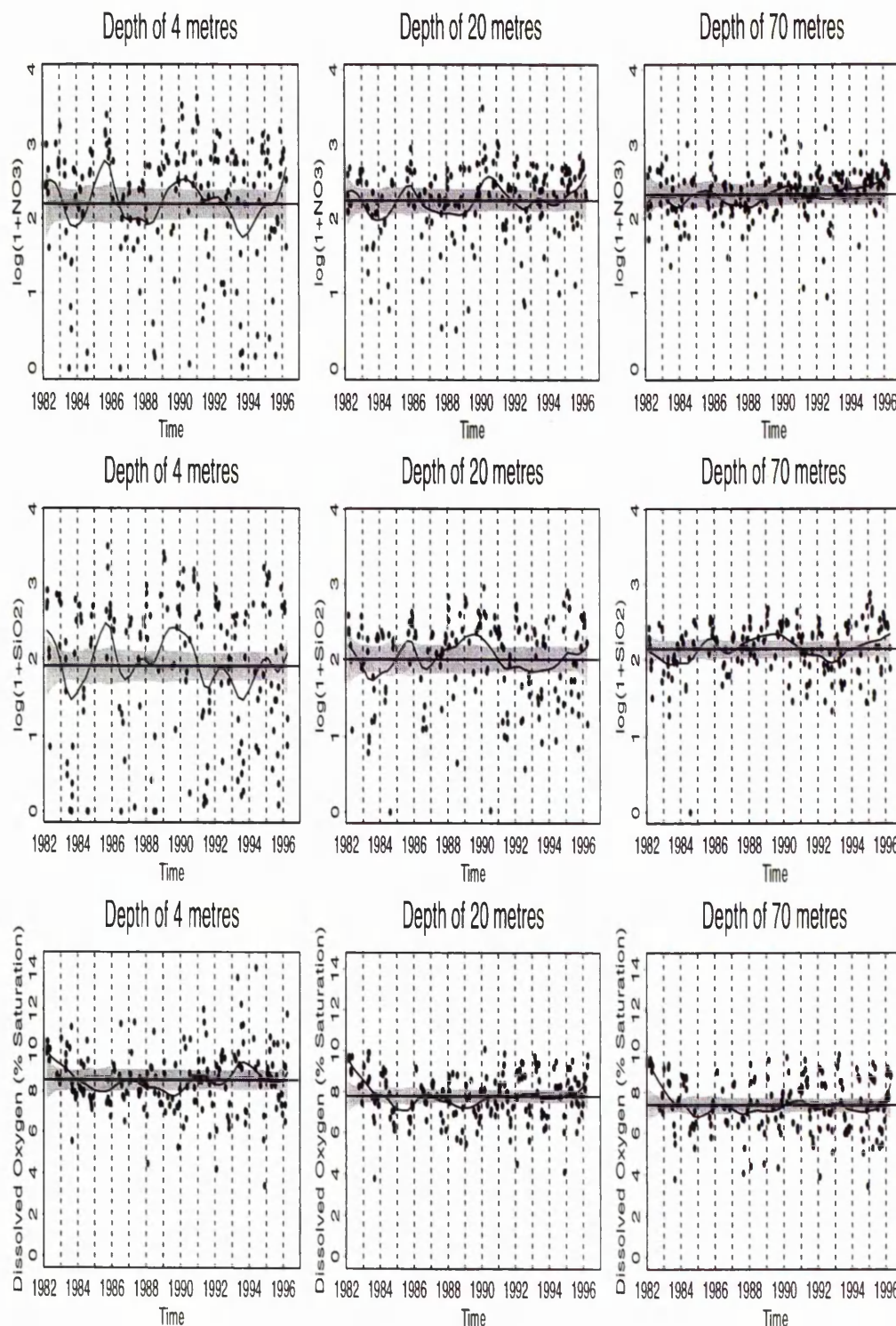
conclusion is supported by comparisons of the smooths with the horizontal 'no-trend' bands. Although there are departures from the bands, the nature of these indicate that there is no long term shift in the mean value. Thirdly, the departures from the reference bands in different years where the sampled dates (and therefore seasons) differ suggests there is a seasonal pattern, especially in surface waters. This aspect of the data is explored further in the following section.

### 6.3.4 Seasonal patterns

The irregular distribution of observations across time also complicates the assessment of the seasonal variation in the variables. This is compounded by the relatively small number of surveys for each season in each year; on average there were 12 surveys annually. One approach to obtaining a crude estimate of seasonal effect is to pool the data from the 14 years into one year. This is accomplished by using the day-of-year value (1,2,...,365) as the 'time' variable. This reflects the assumptions that there is no substantial long-term trend across time for the covariate and that the seasonal fluctuations in variable values are similar within each year. Clearly, each of these assumptions cannot be justified formally but they do allow an ad hoc estimate of seasonality to be constructed.

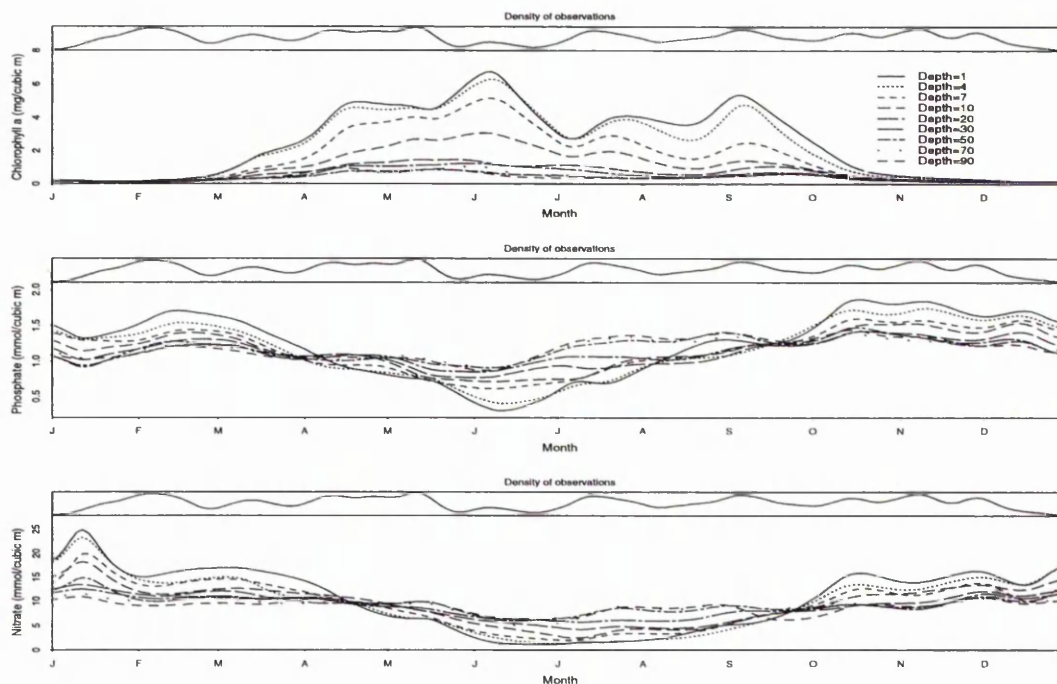
Figure 6.5 shows estimates of the seasonal patterns derived from the collapsed data for three variables: chlorophyll a, nitrate and phosphate (values at the nine depths are shown separately). The estimated trends are generated by a local linear regression applied to subsets of the data consisting of the concentrations of the variables at each of the nine depths. A smoothing parameter of 10 days was used which constructs estimate on each day by 'averaging' observations from the preceding and subsequent 3 weeks. The seasonal (and thus cyclical) nature of the estimated effect was ensured by 'wrapping' the data, i.e. duplicating the observations from January and December to follow and precede the pooled data respectively. This ensures that estimates at the start of the year include 'averaging' values at the end of the year and vice versa. This ensures that the estimated seasonal patterns at the start and the end of the year are continuous.

The upper portion of the plots in Figure 6.5 show a density estimate applied to the 'day-of-year' variable. This gives an indication of the number of values



**Figure 6.4.** Concentrations across time for dissolved oxygen, nitrate and silicate, each at three depths.





**Figure 6.5.** Estimates of seasonal patterns, derived from pooled data, for selected parameters.

averaged to produce the estimates at each day of the year. Thus it can be seen, for instance, that during April and May more values are available for averaging than in, say, December and June when there is a scarcity of observations. This density estimate therefore reflects the survey frequency information contained in Table 6.2. It is important to view the estimated trend in the light of this information since it reflects a level of confidence in the accuracy of the estimated trend (i.e. the more data available the greater the potential accuracy of the estimate).

The pooling and smoothing process used to produce Figure 6.5 has achieved its aim, i.e. the seasonal patterns for chlorophyll a, nitrate and phosphate immediately become obvious. For chlorophyll a, four peaks are observed in April, June, August and September for surface waters concentrations. Care must be taken when interpreting these features however, since the trends are based on a pooled year rather than an actual year. It is possible, for instance, that the peaks



in April and June represent the same seasonal events occurring at different times in different years.

However, these results highlight features in the data which are consistent with observations that Boney (1986) made on the seasonal patterns of chlorophyll *a* in the Firth of Clyde. Boney reported that plankton 'blooms' usually occur in spring and autumn. The smoothed lines also suggest that chlorophyll *a* concentrations rarely reach significant levels at water depths below 10m. This is a reasonable depth for the euphotic zone.

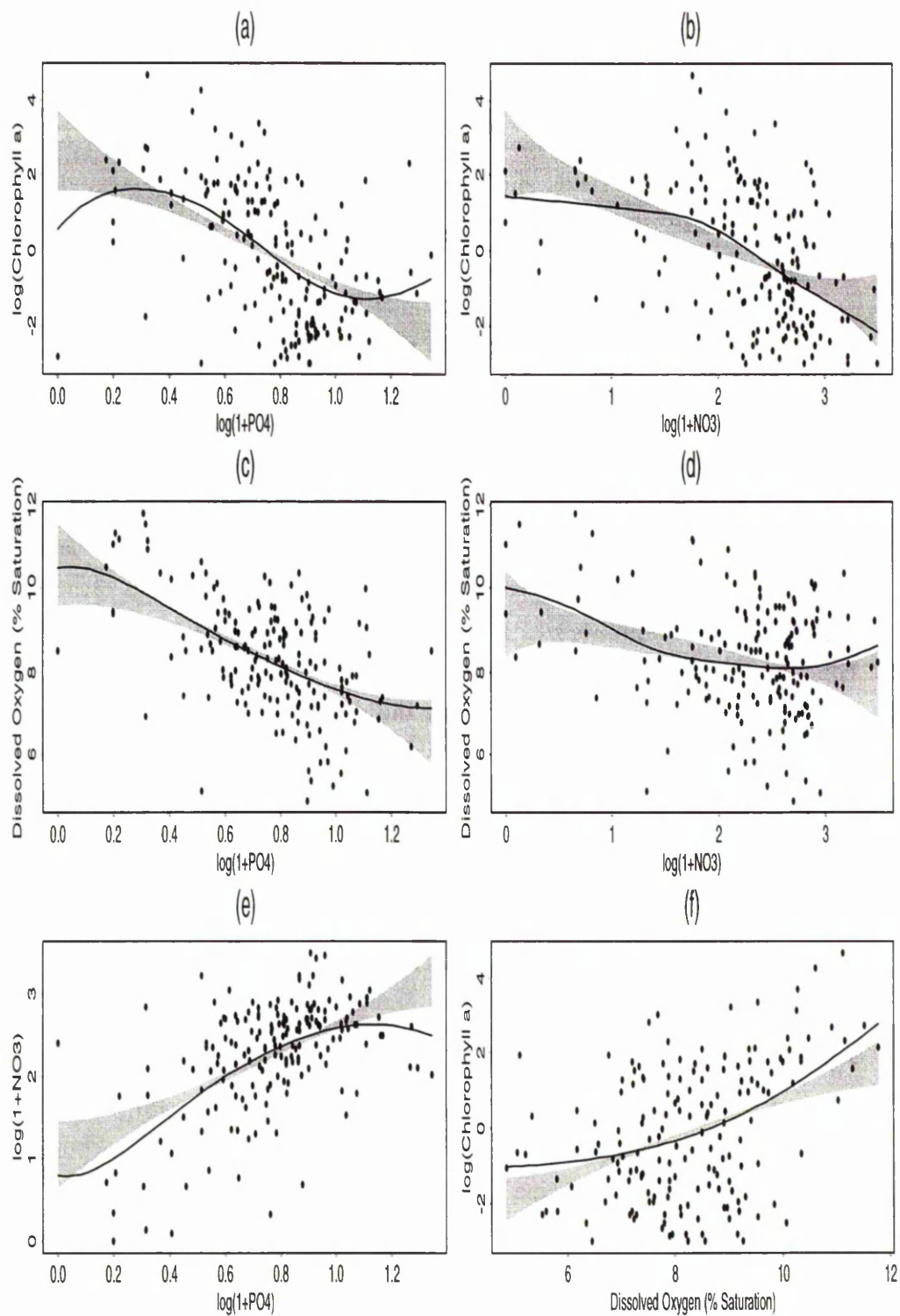
### 6.3.5 Important aspects of the dynamics of the water system

#### Uptake of nutrients by plankton

Given the interest in eutrophication, it is of interest to consider the relationships between the nutrients and plankton growth directly via scatterplots. Figure 6.6 plots six pairs of variables and applies to each a local linear regression to estimate the underlying relationship between the variables. In addition a reference band for a linear relationship is shown.

Plots (a) and (b) in Figure 6.6 display the association between chlorophyll *a* and phosphate and nitrate. There is some evidence of negative correlation between chlorophyll *a* and both phosphate and nitrate. This is most pronounced in the mid range of phosphate values and mid to high range of nitrate values. A similar negative association is shown in the scatterplots of dissolved oxygen (as percent saturation) against phosphate and nitrate (plots (c) and (d) of Figure 6.6). Dissolved oxygen and phosphate show the strongest evidence of a linear relationship whereas dissolved oxygen appears to level off at higher values of nitrate.

Plots (e) and (f) of Figure 6.6 show positive associations between the nutrient compounds and between chlorophyll *a* and dissolved oxygen respectively. In each case there is evidence that at the extremes of the variables the underlying relationship differs from that expected under a linear model.



**Figure 6.6.** Scatterplots of relationships between selected variables, at depth of 7m.

## Stratification

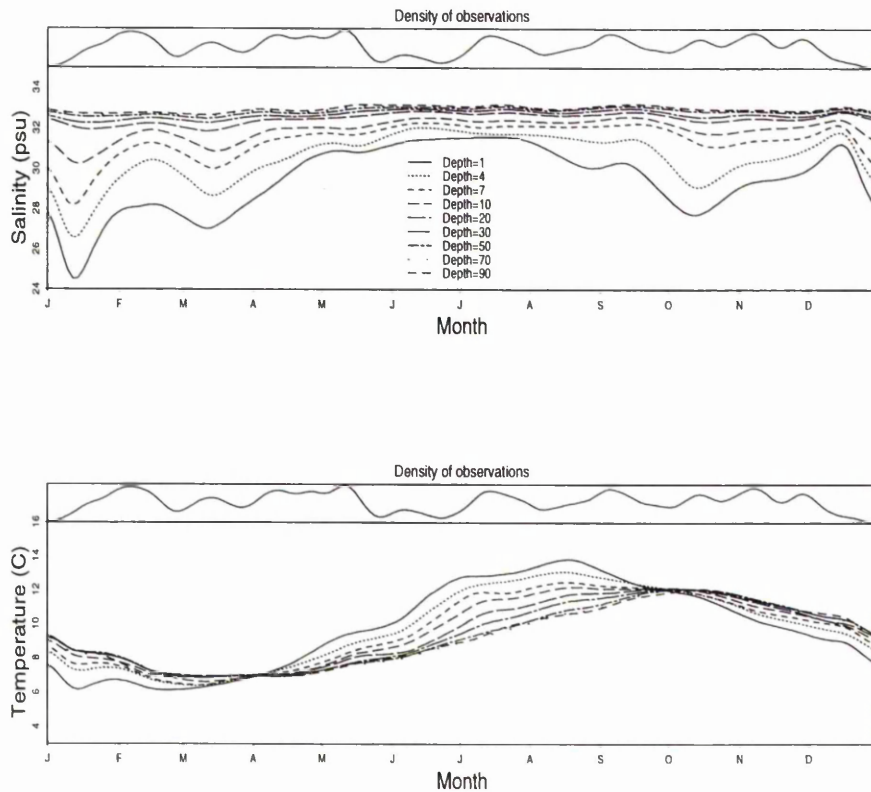
Figure 6.7 shows seasonal plots for temperature and salinity, produced in an identical way to those in Figure 6.5 (see Section 6.3.4). Each of these variables are natural indicators of the degree of stratification in the water column. The estimated seasonal trends show that stratification is more strongly influenced by temperature during the summer months and by salinity during the winter months.

As a means of assessing the seasonal (time) effect at different depths, bivariate nonparametric models were fitted to the data and compared using the methods of Chapter 3. Clearly, if the seasonal effect is consistent across depths then an additive model relating salinity/temperature to the two covariates (day-of-year and depth) would provide a sufficient fit. Alternatively a general bivariate smooth fit would have to be employed. These two model fits were compared using a CFV statistic and a RSS based estimator of  $\sigma^2$ . Given that the two covariates define an irregular grid of points, an undersmoothed version of the bivariate smooth fit was used to estimate  $\sigma^2$ . In both cases (salinity and temperature) the tests conclusively rejected the 'reduced' model (p-values < 0.001). This confirms the patterns shown in Figure 6.7, which indicate that an additive model is not an appropriate description of the observed behaviour. This is a result of the seasonal effect being most pronounced in the surface waters (depth  $\leq 10$ m).

## 6.4 Discussion Points

### 6.4.1 Relationships between variables

It is traditionally assumed that nitrate is limiting in marine waters, although Glibert (1988) states that: 'it is difficult to pinpoint precisely the origin of the concept that coastal and oceanic phytoplankton are nitrogen-limited'. Figure 6.5 can be used to investigate the prevailing assumption that nitrate is limiting in marine waters by comparing the seasonal plots for chlorophyll a, nitrate and phosphate concentrations. Nitrate may be the limiting nutrient for spring phytoplankton growth, since the ratio of nitrate:phosphate is well below 16 (the so-called 'Redfield ratio' (Redfield *et al.*, 1963) which describes the conditions for



**Figure 6.7.** Estimates of seasonal patterns, derived from pooled data, for salinity and temperature.

maximum chlorophyll *a* production). Later in the year, however, phosphate concentrations fall almost to zero, so this may become the limiting factor for autumn plankton growth. The June trough in phosphate closely mirrors the June peak in chlorophyll *a*, suggesting a stronger relationship between these two variables, for surface waters, than between chlorophyll *a* and nitrate. This observation indicates that phosphate may play a more important role in limiting plankton growth in marine waters than is currently assumed.

### 6.4.2 Potential for eutrophication

Furthermore, the Comprehensive Studies Task Team (CSTT, 1994) report indicated that the assessment of whether waters were affected by eutrophication should be based on two criterion. Firstly, whether waters are hypernutrified, with high winter nutrient concentrations of dissolved available inorganic nitrogen (DAIN) significantly above  $12 \text{ mmol/m}^3$ , in the presence of at least  $0.2 \text{ mmol/m}^3$  dissolved available inorganic phosphate (DAIP). Secondly, where this is the case, chlorophyll a concentrations in excess of  $10 \text{ mg/m}^3$  should persist throughout the summer months. Figure 6.5 allows this assessment to be made for the data for the whole 14-year period, 'collapsed' into one typical year.

Winter nitrate concentrations appear to peak in January at about  $25 \text{ mmol/m}^3$  in surface waters, although average concentrations during the winter months of December to February inclusive are lower, at about  $15 \text{ mmol/m}^3$  in surface waters. Corresponding winter phosphate concentrations are about  $1.5 \text{ mmol/m}^3$  in surface waters. Hence, it is concluded that the waters at this site show no evidence of hypernutrification. Secondly, chlorophyll a concentrations do not typically reach  $10 \text{ mg/m}^3$ , so the criteria for algal growth is not met and thus it is furthermore inferred that waters at this site are not subject to eutrophication.

### 6.4.3 Dynamics of the water system

The scatterplots in Figure 6.6 also reflect the relationships amongst the patterns of seasonal trends observed in Figure 6.5. These, it was noted, had peaks in plankton levels coinciding with troughs in the nutrient levels and vice versa. Since %D.O. is also a reflection of the amount of algal activity, it is not surprising that this exhibits a similar relationship with the nutrients as chlorophyll a does. Figure 6.6 shows, though, that the nature of these relationships is not straight forward. Global tests comparing the smooth fits with simple linear fits (as described in Section 1.2.3) each return significant results at the  $\alpha=0.05$  level with the exception of phosphate vs. %D.O., thus confirming the conclusions drawn from the reference bands in Figure 6.6. Keeping in mind that these models are in terms of the transformed variables, the underlying relationships between nutrients and plankton growth are indeed complex.

#### 6.4.4 Effects of stratification

Stratification of the water column is an important feature which determines the degree and extent to which mixing occurs in coastal waters. The degree of stratification is determined by the balance between buoyancy forces which cause stability and stirring effects which cause mixing (Simpson and Rippeth, 1993). Buoyancy forces are generated by freshwater inputs or surface warming, which may lead to stratification if these forces dominate over mixing effects. Mixing forces may be generated by wind stress, internal waves, tides and deep-water replacement (e.g. in sea lochs). Figure 6.7 indicates that temperature effects are the dominant cause of stratification in summer, due to surface water warming. In winter, the dominant effect on stratification is likely to be caused by increased freshwater run-off leading to decreased salinities in surface waters.

### 6.5 Conclusions

This analysis of water quality data collected over a 14-year period at one sampling station in the Firth of Clyde has examined relationships between the variables and changes in the distributions of the variables with depth and through time.

It has been established that median concentrations of most of the measured variables change little in deeper waters at 20-90m depth. The information from one sample collected at, say, 30m depth may therefore be taken as representative of the water column at depth, under conditions of good vertical mixing. A consequent reduction in the number of discrete water samples collected at each 'profile' water sampling station would lead to less time spent at each sampling station. This in turn would allow more stations to be sampled over a wider geographical area in a single survey day, thereby maximising the water quality information gathered using the same resources. This information will also be used in selecting the water depths at which a towed constant monitoring system will be deployed.

The exploratory data analysis did not produce any evidence which suggests long term trends with time for any of the measured variables. This allowed two important conclusions to be drawn: firstly that the two different analytical techniques used during the survey period from 1982 to 1996 produced comparable

results for nutrient concentrations; secondly that the trophic status of the Firth of Clyde waters does not appear to be changing. Comparison of the observed nutrient concentrations with guidelines for the assessment of eutrophication indicates that Firth of Clyde waters are not subject to eutrophication.

The nonparametric modelling tools used produced estimates of seasonal trends for a number of variables which proved consistent with understood biological patterns. The complicated correlation structure of the recorded measurements meant that formal statistical inference was not undertaken, but the exploratory techniques have shed light on the behaviour of different elements of this complicated ecological system.

Lastly, although in environmental terms the data set is one of the most comprehensive available for any coastal water body around the UK, it was necessary to pool the data across the different years in order to assess seasonal variation effectively. This highlights the need for the greater involvement of statisticians at the survey planning stage, so that sample collection can be planned with data analysis in mind.

However, in settings such as environmental modelling, the limitations in resources and a reliance on factors outwith the control of researchers will always be a barrier to the translation of a statistically robust sampling plan into a practical monitoring programme. Hence the methods used in this analysis, particularly those involving nonparametric modelling, offer some measure of a solution by extracting from the data information which can be used to better plan and assess the monitoring of this complex system.

## Chapter 7

# Assessing long term Cyclosporin treatment effects in kidney transplantation outcomes.

### 7.1 Introduction and Background

This chapter presents the results of a retrospective analysis of data describing the long term outcomes of a group of kidney transplant patients. Of particular interest is the relationship between the condition of the transplanted kidney, and the pattern of early treatment with the immunosuppressive drug *cyclosporin*.

Cyclosporin is widely and routinely prescribed to patients following kidney transplantation. The short term effects of the drug are well understood, namely it prevents acute rejection of the new kidney by interfering with the normal activation of the immune system. Unfortunately these improvements do not carry over to long term graft survival when compared with previous immunosuppressive treatments (see Morris *et al.* 1987). Furthermore, in addition to the considerable cost of treatment, prolonged exposure to the drug may result in infections and tumours and, in addition, cyclosporin itself may cause kidney damage, i.e. nephrotoxicity (Myers *et al.* 1991).

In light of these factors, the usual clinical practice is for large doses proportional to the patient's weight to be administered soon after transplantation.



These doses are gradually reduced over a period of about one year, after which the dose of cyclosporin may remain constant or be subject to small alterations according to the presence or absence of side effects and the absorption of the drug as revealed in trough blood levels of cyclosporin.

With regard to the long term outcome, even kidney transplants which are functioning well at one year are unlikely to continue to do so indefinitely, and ultimately most will deteriorate slowly and eventually fail. This process, known as progressive graft dysfunction (PGD), is usually due to a slow form of rejection (chronic rejection) but may be contributed to by other factors such as cyclosporin nephrotoxicity.

Published studies have investigated the long term effects of cyclosporin by making comparisons between randomly allocated groups which differ in that cyclosporin is withdrawn from one and replaced by an alternative treatment at periods ranging from 1-12 months following transplantation. See, for example, Hollander *et al.* (1995), Hall *et al.* (1988) & Rowe *et al.* (1994). In general these studies found an increased incidence of acute rejection in the few months following conversion but this seldom led to a difference in long term outcome when compared to the group with continued treatment. This suggests that prevention of chronic rejection by cyclosporin is to some extent nullified by its nephrotoxicity. As a result of this there is some uncertainty about the optimum long term cyclosporin dose following kidney transplant.

This analysis aims to contribute to the understanding of the long term effects of cyclosporin. The study involves a retrospective analysis of data from patients who have received a prolonged treatment with cyclosporin. It differs from the studies mentioned above, firstly on the grounds that it is not a randomised controlled trial and secondly in that it is investigating the effect of different intensities and patterns of cyclosporin treatment. This requires measures reflecting characteristics of cyclosporin treatment over a period of time and it is here we employ nonparametric modelling techniques to aid in the analysis.

## 7.2 Previous and Current Work

This work builds on the results of earlier collaborative work between Eileen Wright, previously of the Department of Statistics at the University of Glasgow, and Drs. Peter Rowe and Maureen Lafferty of the Renal Unit at Glasgow's Western Infirmary (Wright, 1995). The data amassed in the course of that work have been revisited as part of this analysis, and their results provided a very useful starting point.

This earlier work focused on the identification of, and characteristics associated with, the onset of Progressive Graft Dysfunction (PGD). The commencement of PGD was detected via an algorithm which applied simple local linear methods to traces of serum creatinine levels in order to detect steady increases. Linear increases in the transformed (negative reciprocal), body-weight-adjusted serum creatinine levels are well recognised as indicating deterioration of the kidney. The algorithm was applied to 446 individual grafts resulting in a classification of the outcome to date as one of the four categories:

- ◇ stable, functioning well;
- ◇ stable, functioning poorly;
- ◇ progressive graft dysfunction (PGD), and
- ◇ acute graft loss (AGL).

Defining a PGD classification as an 'event' and the remaining classifications as 'censored' observations, survival analysis was used to assess the effect of a range of clinical covariates, including cyclosporin dosage and blood levels, on kidney outcome. Of these, the occurrence of rejection episodes, the source of the kidney (live donor or cadaver) and the history of blood pressure appeared as significant variables in proportional hazard models applied to the data. In contrast to previous studies by Mickey *et al.* (1990), Gjertson (1991) and Salomon (1992), however, factors such as the age of the donor and recipient, the degree of graft match and, most importantly, the levels of cyclosporin dosage were not found to be significant influences on the onset of PGD.

The analysis presented in this chapter, conducted in collaboration with Dr Stuart Rodger of the Western Infirmary's Renal Unit, differs in a number of important respects from this earlier work.

**Definition of Outcome.** The outcome of interest in this current work is the long term status of the transplanted kidney. Unlike earlier work where an 'event' was defined as a specific form of deterioration (PGD) this present work defines all departures from stability as 'events'. Furthermore, this present work doesn't employ an algorithm to determine outcome, but rather uses expert judgement to assign patients to the outcome categories.

**Population of Interest.** Whereas all transplants were considered in the earlier analysis, the current analysis is confined to those patients who had a positive short term outcome, defined as a stable kidney at two years post-transplantation. Thus, by considering this cohort, the outcome of interest becomes *long term* outcome.

**Modelling of cyclosporin.** A consequence of the changed outcome focus is that cyclosporin must be modelled differently. Since it is the treatment pattern during the initial period post transplant which is of interest, summary measures of cyclosporin dosage and blood levels are derived from the first two years post-transplantation and modelled as covariates (*not* time dependent).

**Role of other covariates.** Whereas a range of covariates were considered previously in their own right, the current study includes a baseline set of covariates, which attempt to account for known and suspected influential factors identified in previous studies. Although a greater understanding of the role of these variables will be gleaned this is of secondary importance.

In summary, the primary aim of the current analysis is to provide some evidence toward answering the question: *In those patients whose short term experience is favourable, is the cyclosporin treatment over the initial period post-transplantation a significant factor in determining the condition of the kidney in future years?*

## 7.3 Study Design

### 7.3.1 Selecting a cohort of patients

One of the key aspects of this analysis is the cohort of patients whose records were analysed. Patients were included in the analysis according to the following criteria:

1. Patient received a kidney transplant through the Renal Unit at Glasgow's Western Infirmary during the period January 1984 to December 1990.
2. Two years after the transplant date the patient had a stable, functioning kidney.
3. The patient's cyclosporin treatment was not terminated after one year as part of the *Cyclosporin Conversion Study*.

The period in which the transplants occurred (1984-1990) is suited to this study since it is sufficiently removed from the present to ensure that long term effects can be observed. This criterion also coincides with the work of Wright (1995) where data describing the post-transplantation treatment and condition of kidney transplant patients seen at the Western Infirmary Renal Unit<sup>1</sup> between January 1984 and December 1990 formed the basis of Wright's study.

The second criterion is also related to the objective of examining long term effects. By only selecting patients whose kidneys perform well at the two year mark, time is allowed for the patient's condition to stabilise and hence short term, temporary deterioration is overlooked. Both patients who experience kidney failure and those who die (any cause) during the two years following the transplant are not included in this study. The second criterion does not, however, exclude patients who experienced a complication shortly after the transplant but regained kidney function within two years of the transplant. This situation may indeed be a contributing factor to the long term condition of the kidney, but it was not a basis for excluding patients from the analysis in the first instance.

---

<sup>1</sup>Patients were referred from a number of sources, namely GRI, STOB, WIG, RAIGMORE, DUMFRIES, RHSC

Of the 370 patients satisfying the first criterion, Wright's classifications identified 255 transplants as not experiencing failure in the first two years. A further 69 transplants were classified as 'stable' using data spanning less than two years since transplantation. Each of these 324 patients' records at the two year mark was reviewed and a clinical rule was used to verify that they met the second criterion for inclusion in the study. Namely, a patient's kidney was classified as stable at the two year mark if:

- ◇ their creatinine level at the two year mark was less than  $150 \mu\text{mol/l}$  or
- ◇ the increase in creatinine from 1 year to 2 year post-transplantation was less than 30%.

Using these criteria together with the results of Wright's algorithm, 292 patients were identified as being suitable for inclusion in the analysis.

The final requirement for inclusion ensures that the patient's cyclosporin experience was not interrupted artificially. The *Cyclosporin Conversion Study* was a randomised trial to assess the effectiveness of alternative treatments to cyclosporin. As such, a randomly selected group of patients stopped receiving cyclosporin approximately one year after transplantation. Indeed, it may be of interest to compare the long term outcome of these patients with those who continued with their cyclosporin treatment, but the focus will be on the group who received a standard cyclosporin treatment.

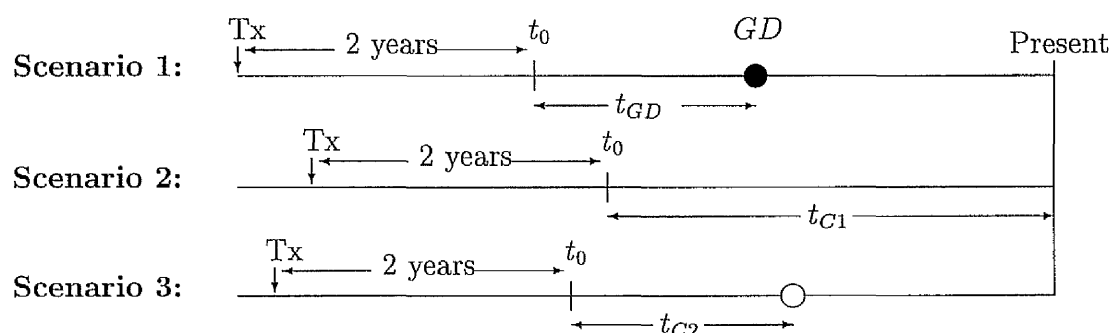
There were 217 patients who satisfied all three criteria.

### 7.3.2 Defining and determining outcome

In addition to Wright's data and results, data describing the outcome of interest and the cyclosporin treatment during the first two years post-transplantation are required. The former data are described here, cyclosporin measures are discussed in Section 7.4.

Recall that the outcome of interest is the condition (stable/unstable) of the transplanted kidney at the most recent observation. In addition to this classification, the *time to event* is also required. The following diagram illustrates three different meanings of *time to event* depending on the observed behaviour of the

kidney. In each case 'Tx' represents the transplantation and ' $t_0$ ' is the date two years post-transplantation,



Scenario 1 represents a patient who was observed to experience graft dysfunction (GD) leading to failure, the commencement of which is marked by  $\bullet$ . Associated with this 'event' is the time since  $t_0$ , namely  $t_{GD}$ . Scenario 2 shows a patient whose kidney function is observed as stable through to the most recent observation ('Present'). Note that the associated time to event,  $t_{C1}$ , is censored. That is,  $t_{C1}$  is a lower bound for the time to GD although we are not to know if it ever would occur. Scenario 3 also describes a patient who was not observed to experience GD but for some reason, e.g. death unrelated to kidney function, left the study (this event is represented by  $\circ$ ). The time to this event,  $t_{C2}$ , is also a censored time since we did not observe the long term outcome of the kidney.

Two variables are therefore necessary to convey the two features of the outcome. The first is whether or not the kidney was last observed as stable (binary variable). The second is an associated time (continuous variable), recording either *time to commencement of graft dysfunction* or *time to censoring*.

These data were obtained via an expert review of individual patients' records. In the first instance, patients were classified into one of six categories which describe the outcome in considerable detail, namely

1. Functioning well;
2. Functioning at time of death (unrelated to kidney performance);

3. Functioning yet experiencing PGD;
4. Functioning yet experiencing PGD at time of death;
5. Failed after a PGD;
6. Failed after an Acute Graft Dysfunction<sup>2</sup>.

Relating these classifications to the binary outcome necessary for the analysis, the first two categories define the censored (i.e. stable) observations and the remainder are the observed 'events' (i.e. deteriorating or failed). Note that patients *lost to follow up* pre-deterioration are included in the first category.

A similar criterion to that used at the two year mark was employed to define if and when a patient's kidney began to deteriorate, namely creatinine greater than  $150\mu\text{mol/l}$  and greater than a 30% increase on the second year creatinine level.

Table 7.1 shows the breakdown of *outcome*<sup>3</sup> for the 217 patients suitable for analysis. It shows that 64% of patients had stable functioning kidneys at the last observation. Of the 36% whose kidney function had deteriorated 92% experienced progressive graft dysfunction rather than acute graft dysfunction.

Outcome	1	2	3	4	5	6	missing
Frequency	118	21	27	3	39	6	3

**Table 7.1.** Number of patients corresponding to each level of the *outcome* variable

### 7.3.3 Demographic and clinical variables

A number of variables describing the patients' demographic and clinical factors at the time of transplant were available. Although not all of these are considered in the present analysis, they are listed here for completeness:

<sup>2</sup>An *Acute Graft Dysfunction* is defined as a deterioration in renal function leading to graft failure within a 6 month period.

<sup>3</sup>The detailed code for three patients was not available, but their binary outcomes were: 1 stable, 2 unstable

- ◇ sex of patient
- ◇ age of patient at time of transplant
- ◇ diabetes mellitus (yes, no)
- ◇ diagnosis edta code
- ◇ hospital source (GRI, STOB, WIG, RAIGMORE, DUMFRIES, RHSC)
- ◇ donor's age at time of transplant
- ◇ donor's sex
- ◇ source of donated kidney (local, import)
- ◇ type of dialysis (haemo, capd, mixed)
- ◇ type of transplant (cadaver, live)
- ◇ information on types A, B, DR of mismatches
- ◇ panel reactive antibodies pre-transplant
- ◇ peak panel reactive antibodies
- ◇ number of transplants prior to the transplant recorded in the data
- ◇ commencement dates of rejection episodes (up to three)
- ◇ number of rejection episodes treated with steroids
- ◇ number of rejection episodes treated with OKT3
- ◇ number of rejection episodes treated with ATG
- ◇ time on dialysis until transplant
- ◇ time between transplant and kidney's first functioning
- ◇ creatinine level at 3 months post-transplantation



Of these variables, *sex*, *patient's age*, *donor's age*, *type of transplant*, *transplant number*, *3 month creatinine level* and *EDTA code*<sup>4</sup> are to be considered in the analysis. In addition, *the number of rejection episodes in the first two years post-transplantation* and a *mismatch category* were calculated for inclusion in the analysis. Three mismatch categories were defined as follows:

1. No mismatches of any type
2.   ◇ 1 mismatch of type A, none of B, none of DR
  - ◇ 0 mismatches of type A, 1 of B, none of DR
  - ◇ 1 mismatch of type A, 1 of B, none of DR
3. At least two mismatches of type A or B, non-zero number of type DR

## 7.4 Cyclosporin Variables

In this section descriptive and nonparametric modelling methods are used to derive measurements which describe each patient's *cyclosporin treatment* over the two years immediately post-transplantation. These measures are then used to model the long term outcome using proportional hazard models in Section 7.5. Hence the present task of deriving cyclosporin measures involves a somewhat novel use of nonparametric modelling techniques since they are not here used to capture the features of a *response* variable, but rather used to define *covariate* measures. To accomplish this, several of the nonparametric techniques introduced and discussed in previous chapters are employed.

Data describing the cyclosporin treatment over the first two years after transplantation were obtained directly from hospital records. This information was available in two forms; the *doses prescribed (mg)* and the trough *blood levels*. In addition, weight data were also available which allowed the 'raw' doses prescribed to be converted to *dose per kilo weight* which reflect the clinical practice of prescribing weight adjusted doses.

The measures derived from these data seek to convey three intuitive features:

---

<sup>4</sup>EDTA code has been collapsed to two categories: "80"/"81" (diabetes) and "others"

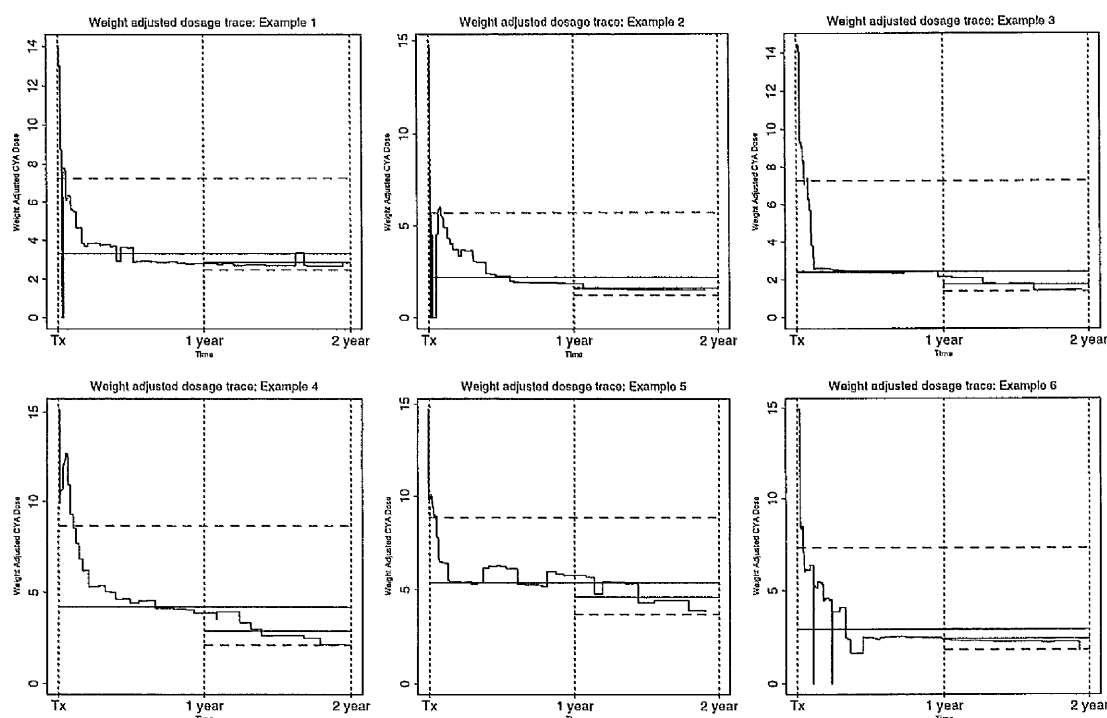
- ◇ the *magnitude of treatment*;
- ◇ the *variation in the treatment* and
- ◇ the *pattern of the treatment*.

The *magnitude* of cyclosporin exposure is of interest because of the connection between deterioration in the kidney function (PGD) and cyclosporin nephrotoxicity which was highlighted in the introduction. The *variability* within the values may also be an important factor influencing subsequent kidney performance since a highly variable exposure may induce instability in the kidney performance. Finally, the general *pattern* or trend in the treatment over the two years is classified by noting whether the cyclosporin treatment followed a decreasing trend or whether there were substantial increases in the cyclosporin exposure during the two years.

Of the 217 patients, 206 had adequate dose history data, 205 had adequate weight information to adjust the dose histories and 140 patients had blood levels data. Figures 7.1 and 7.2 show weight adjusted doses and blood levels of cyclosporin for the same six patients. These plots highlight the differences between dose data and the levels data. The doses are prescribed and maintained at constant levels until the next change (as indicated by the ‘step’ nature of the plot) and as such are free of inherent variability. The blood level readings, in contrast, are irregularly spaced samples over time and have inbuilt variation reflecting biochemical rhythms in blood composition. While the dosage data is a precise record of the cyclosporin prescribed, the levels contain within them variability which masks the true underlying trend. Given these differences, the two forms will be analysed differently to derive measures of interest.

#### 7.4.1 Dosage data

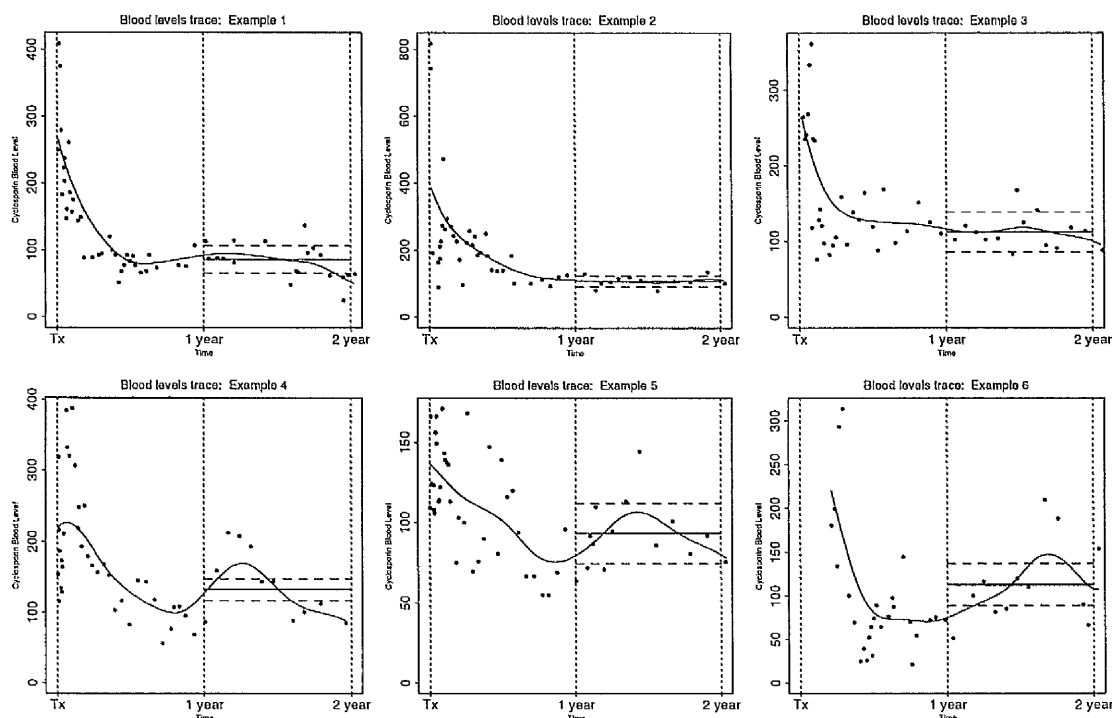
The dosage measures are calculated from both the raw and weight adjusted data over two periods of time: the full two years post-transplantation and the second year post-transplantation. The reason for considering the second year on its own is that it ignores the initial very high doses which are prescribed immediately after the transplant.



**Figure 7.1.** Traces of the cyclosporin weight adjusted dose data for six patients. Also shown are the measures of magnitude and variation over the two years and the second year only.

The following measures, corresponding to the *magnitude*, *variability* and *pattern* of exposure have been calculated:

- ◇ Time weighted average of both the raw and the weight adjusted cyclosporin doses. This corresponds to the area under the step curves in the dosage plots of Figure 7.1 divided by the duration of interest (365 or 730 days).
- ◇ Standard deviation of the raw and adjusted doses around the time weighted average.
- ◇ Binary variable signalling if a patient's cyclosporin doses followed an approximately monotonically decreasing pattern. This is coded 1 ('YES') if no consecutive doses increase by more than  $25mg/0.35mg(kg)^{-1}$  for the raw/weight adjusted doses. These thresholds represent clinically significant changes in treatment.



**Figure 7.2.** Traces of the cyclosporin levels data for six patients with an estimate of the underlying trend using local linear regression (smoothing parameter of 60 days). An estimate of the average level over the second year is shown together with  $\pm$  the estimate of the standard deviation of the levels around the underlying trend.

Figure 7.1 displays the weighted average (solid line) minus one standard deviation (dotted line) over the second year post-transplantation and the weighted average (solid line) plus one standard deviation (dotted line) over the full two years post-transplantation for these weight adjusted dose data.

#### 7.4.2 Blood levels data

Unlike the doses, the levels data do not provide a complete description of the patients' level histories. As mentioned above, the levels are available for a sample of days throughout the two year period. These usually correspond to visits to an outpatient's clinic and therefore exhibit irregular intervals between observations and varying frequencies throughout the period. This is reminiscent of the

challenges encountered with the environmental time series considered in Chapter Six.

Another complication in the levels data concerns the values themselves. The levels of cyclosporin recorded in the blood samples is related to a number of unknown factors such as the time of day, time of last meal, time of last intake of cyclosporin etc. Although guidelines are in place which attempt to introduce uniformity to the level readings the inherent measurement error must be taken into account when variables are derived.

Although these complexities are associated with the blood levels data, there are also reasons for preferring measures of cyclosporin exposure derived from the levels rather than the doses. The blood levels of cyclosporin are a direct measure of the patient's intake of cyclosporin. Different absorption rates are therefore reflected in the levels data whereas these are missed by the dose data. Also, patients self-administer the prescribe doses for most of this two year period, therefore the question of compliance arises.

For these reasons the blood level traces do not share a common trend. For instance, in Figure 7.2 the first row of patients' levels show a marked decrease in the early months and then settle down to a modest, fairly constant level in the second year. The second row of patients' levels shows a similar decrease in the first year, but the second year's pattern is characterised by an increase in levels followed by another decrease.

Given these properties of the blood levels data a nonparametric modelling approach has been used to derive measures of exposure. This has taken the form of direct estimation of the underlying trend and its first derivative by local linear regression and also the estimation of the variation of the recorded levels around the unknown trend using difference based estimators (Gasser *et al.* (1986)). Both of these techniques have been described in detail in previous chapters. The first derivative of the underlying curve is estimated simply by the slope parameter estimate  $\hat{\beta}_1$  of the locally weighted linear fit defined by Equation 3.1.

The use of local linear regression requires the specification of a bandwidth to govern the amount of smoothing. Since the same scale (days) applies in each case it is attractive to use the same smoothing parameter for each patient, regardless of the distribution of days at which levels are available. This ensures that the

same degree of smoothing is employed, in the sense that data from identical time spans are used to estimate the underlying trend for each patient. A range of smoothing parameters were considered and  $h = 60$  days was found to pick up sufficient detail without being too 'bumpy'.

Because of the inbuilt variability and the added instability immediately after the transplantation, only measures from the second year post-transplantation are derived. Furthermore, only patients whose levels approximately span the second year are included. This reduces the number of patients with sufficient blood levels from 140 to 120. For each of these patients the following measures were derived:

- ◇ Average of the estimates of the underlying levels' trend at 'evaluation dates' selected to span the available data at regular (7 day) intervals. This is an estimate of the area under the smooths over the second year divided by the exact span (in days) of the levels data over this second year.
- ◇ Square root of the difference based estimate of the variation in the blood levels around the underlying trend using the levels falling in the second year only. This is the estimator which was extended in Chapter Two for use in a bivariate setting. There it was shown that a difference based estimator demonstrated attractive bias properties by estimating the underlying trend locally but without requiring the specification of a smoothing parameter.
- ◇ Average of the estimates of the rate of change of the underlying trend at the evaluation dates. This measure of the general pattern of the cyclosporin levels is chosen rather than a binary variable reflecting monotonicity (as used for the doses) since the complexities of the data make it unwise to place too much emphasis on a single estimate of the underlying trends properties. This measure can be interpreted to be the overall 'direction' of the trend in the levels over the second year.

The levels plots in Figure 7.2 display the estimated underlying trend and the average estimated underlying trend  $\pm$  one estimated standard deviation of the responses around the estimated trend in the second year.

Variable	Period	N	Median	Mean	Min.	Max.
Avg. Wtd. Dose	Both Years	205	3.548	3.66	0.907	8.84
Avg. Wtd. Dose	2nd Year	205	2.776	2.79	0.000	5.29
St. Dev. Wtd. Dose	Both Years	205	4.781	4.80	1.473	9.68
St. Dev. Wtd. Dose	2nd Year	205	0.410	0.51	0.000	3.21
Avg. Raw Dose	Both Years	206	239.5	247.2	86.2	701.9
Avg. Raw Dose	2nd Year	206	200.0	193.1	0.00	428.5
St. Dev. Raw Dose	Both Years	206	306.0	311.1	93.7	687.6
St. Dev. Raw Dose	2nd Year	206	25.3	34.64	0.00	436.1
Avg. Lev. Trend	2nd Year	120	95.97	97.62	44.71	195.10
Avg. Lev. Rate of Chg.	2nd Year	120	-0.044	-0.035	-0.481	0.477
St. Dev. Lev.	2nd Year	120	20.76	31.38	5.01	223.50

**Table 7.2.** Continuous cyclosporin covariates together with their descriptive statistics.

Variable	Period	N	Yes	No
Mono. Wtd. Dose	Both Years	205	64	141
Mono. Wtd. Dose	2nd Year	205	170	35
Mono. Raw Dose	Both Years	206	60	146
Mono. Raw Dose	2nd Year	206	160	46

**Table 7.3.** Binary cyclosporin covariates indicating the monotonicity of doses.

### 7.4.3 Summary of cyclosporin variables

In this section, nine variables have been defined, each of which reflects some aspect of the cyclosporin experience of each patient. Table 7.2 lists the continuous variables along with some descriptive statistics. Table 7.3 shows the distribution of the binary variables which indicate whether both forms of the dose data are constantly decreasing over both the two year period and the second year only.

## 7.5 Analysis and Results

Having defined the variables of interest in the preceding sections, we are now in a position to analyse the data. Given the occurrence of censoring in the response

variable, the methods used are taken from the survival analysis literature (see Cox and Oakes (1984)). Particular use will be made of the *proportional hazards model* (Cox, 1972) which is by far the most commonly used approach. The advantage of such an approach is that it has a structure which can incorporate continuous explanatory variables while retaining a flexibility which permits an undefined (nonparametric) underlying hazard function for the data.

Section 7.5.1 considers the covariates of interest individually and Section 7.5.2 considers their effects in the presence of other covariates.

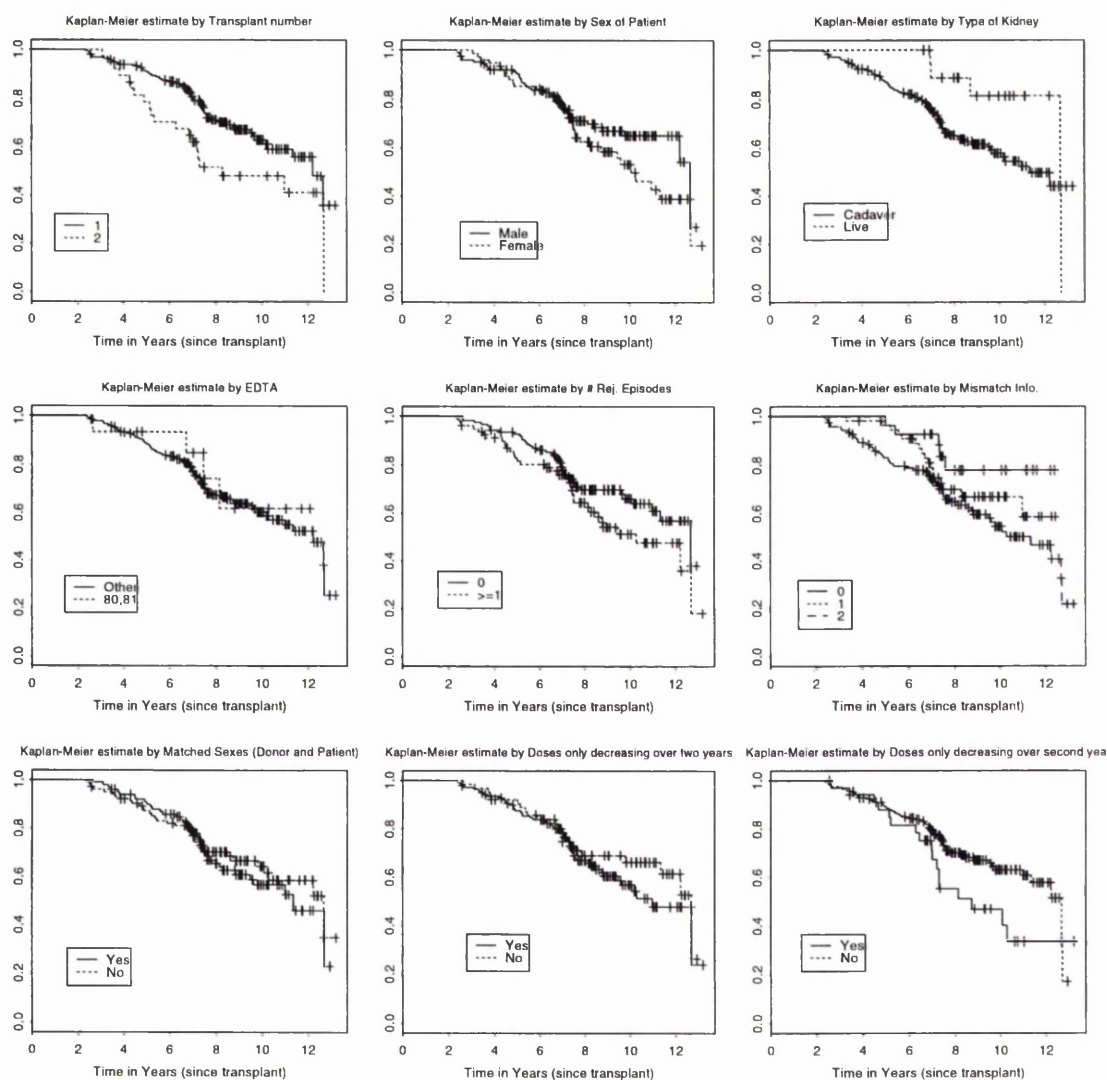
### 7.5.1 Univariate analysis

Although interest ultimately lies in the combined effect of the variables of interest, it is informative to consider the apparent effects on kidney outcome of each covariate individually. Figure 7.3 shows Kaplan-Meier estimates of the survival function estimate at the different levels of the categorical factors using the 206 patients with sufficient raw dose information.

Table 7.4 lists the categorical covariates and the p-value from a Mantel-Haenszel (log-rank) test under the null hypothesis that the survival functions associated with different levels of the individual variables are equivalent. Since this current analysis does not involve cyclosporin, clearly we use all the data available (i.e.  $n = 206$ ). To aid comparisons and the interpretation of later analyses involving the cyclosporin levels variables, results based on the subset of patients with sufficient blood levels ( $n = 120$ ) are presented alongside the results based on all the data.

The results of Table 7.4 indicate that *transplant number* is a key factor in determining the long term outcome of the transplanted kidney. Variables for which there is some evidence of an association with long term outcome include *number of rejection episodes*, *mismatch group*, *type of kidney* and *monotone dose patterns in the second year*. The sex based covariates and the EDTA grouping show no sign of an association. Given the reduced sample size of the 'levels' data, and thus the reduced power of these tests, it is not surprising that with the exception of *transplant number* none of the other variables showed signs of differences between the variable levels.





**Figure 7.3.** Kaplan-Meier estimates of the survival function estimate at the different levels of the categorical factors based on the 206 'raw dose' data.

Covariate	P values from log-rank tests	
	Raw Doses n=206	Blood Levels n=120
Transplant number	0.021	0.007
Sex of Patient	0.165	0.309
Type of Kidney	0.094	0.387
EDTA Group	0.661	0.800
No. Rej. Episodes	0.099	0.642
Mismatch Group	0.069	0.435
Matched Sexes	0.553	0.842
Monotone Dose (Both years)	0.36	-
Monotone Dose (2nd only)	0.07	-

**Table 7.4.** Results of log-rank tests examining the effects of categorical covariates on long term kidney outcome

Interest now turns to the continuous variables and the effect these have on the outcomes of patients. This requires a modelling approach which incorporates and examines the effects of one or more covariates on the survival times. Here we employ the proportional hazards model which models the hazard function at time  $t$  of patient ' $i$ ' as:

$$h_i(t) = h_0(t) \exp(\beta^T \mathbf{x}_i),$$

where  $h_0(t)$  is an (unknown) baseline hazard function,  $\mathbf{x}_i$  is the vector of covariate values associated with the  $i$ th patient and  $\beta$  is a vector of unknown parameter values. Interest lies primarily in the estimation of these parameter values and using them to perform inference on the covariate effects.

Table 7.5 lists the continuous covariates of interest and the key results when each covariate is included solely in a proportional hazards model. These results indicate that the *age of the donor*, the *creatinine levels three months after the transplant* and the *number of rejection episodes* may each have an influence on kidney outcome, although the magnitude of this effect may be quite small. Of the cyclosporin measures considered individually, the *standard deviation in the unadjusted doses over the two years*, *standard deviation of the levels around the underlying trend over the second year* and the *approximate monotonicity of the*

Covariates	Raw Dose		Adj. Dose		Bld Lev.	
	$e^{\beta}$	P-val	$e^{\beta}$	P-val	$e^{\beta}$	P-val
Age of Patient	0.99	0.47	0.99	0.47	0.99	0.38
Age of Donor	1.02	0.03	1.02	0.03	1.01	0.14
Creatinine at 3 mth	1.00	0.04	1.00	0.04	1.00	0.06
No. of rejection episodes	1.46	0.02	1.46	0.02	1.20	0.46
Magnitude (two years)	0.998	0.26	0.994	0.95	-	-
Magnitude (2nd year)	0.999	0.78	1.060	0.58	1.00	0.50
Variability (two years)	0.997	0.04	0.824	0.09	-	-
Variability (2nd year)	0.997	0.32	0.817	0.30	1.01	0.02
Pattern (two years)	0.636	0.10	0.791	0.36	-	-
Pattern (2nd year)	0.576	0.03	0.605	0.07	0.86	0.90

**Table 7.5.** Continuous covariate results from univariate proportional hazard model fits

*raw doses in the second year* show the most evidence of statistical significance in the models.

### 7.5.2 Multivariate analysis

Although the results of the previous subsection are useful in assessing the influence of individual variables, real interest lies in the way these variables combine to influence the patients' outcomes. This leads us to consider proportional hazards models with several covariates.

The main aim is to assess the significance of the effect of variables reflecting the cyclosporin history on each patient's long term outcome. It is important, however, to ensure that other factors known to affect long term outcome are taken into account. Based on expert knowledge and results of previous studies (Morris (1994) and Woo *et al.* (1999)), nine variables were identified (see Section 7.3.3) as of sufficient importance to be included in the model regardless of their statistical significance.

The proportional hazards model with these nine variables will be referred to as the *baseline* model. The details of the fitted baseline model are shown in Table 7.6, fitted to different subsets of patients associated with each of the

cyclosporin history measures used<sup>5</sup>. When these variables are modelled together, we see that the transplant number and donor age show signs of significance in both the  $n = 205$  and  $n = 120$  sets.

Covariates	Adj. Doses		Blood Levels	
	$e^\beta$	P-value	$e^\beta$	P-value
patient's sex	1.45	0.14	1.93	0.07
transplant number	2.35	0.00	3.56	0.00
transplant type	0.33	0.03	0.37	0.14
EDTA group	1.35	0.57	2.71	0.22
No. of rejection episodes	1.14	0.49	1.02	0.96
Mismatch grouping (contrast 1)	1.28	0.35	0.86	0.62
Mismatch grouping (contrast 2)	1.34	0.01	1.20	0.19
Age of patient	0.99	0.71	0.99	0.35
Age of donor	1.02	0.00	1.03	0.01
Creatinine level at 3 month	1.00	0.37	1.00	0.13

**Table 7.6.** Details of the 'baseline' fit

In determining whether the addition of the cyclosporin variables to the baseline model significantly improve the fit, care must be taken. The estimates of the covariate effects from these multivariate models are not strictly independent of other covariates present in the baseline. Tests of significance based on these estimates and their standard errors (Wald tests) are consequently prone to misinterpretation. A safer approach is to compare nested models via the log likelihood ratio test and assess the contribution of covariates by reference to a  $\chi^2$  distribution.

Table 7.7 shows the results of fitting a number of models constructed by adding the cyclosporin variables to the baseline model. In each instance, the log likelihood ratio of the model with the cyclosporin measure(s) to the baseline model is shown together with its associated p-value. In addition to adding the individual measures of cyclosporin, the interaction between the *magnitude* and

<sup>5</sup>Although there is one extra patient in *unadjusted doses* subset than the *weight adjusted doses* subset, only results pertaining to the latter are listed since the model fits are approximately equivalent

*variability* measures is explored by including the product term in a model with the individual terms and comparing this to the baseline.

The results show that the strongest relationship is associated with the *variability* measures and this is true across the three types of cyclosporin data considered. However, these model comparisons indicate that none of the cyclosporin related measures appear to make a significant contribution to the baseline model.

Model	Unadj. Doses		Wt Adj. Doses		Blood Levels	
	LLRT	P-value	LLRT	P-value	LLRT	P-value
Baseline	32.2	-	32.2	-	23.4	-
+Magnitude (two years)	33.0	0.37	32.6	0.53	-	-
+Magnitude (2nd year)	32.2	1.00	32.3	0.75	23.6	0.65
+Variability (two years)	34.7	0.11	34.8	0.11	-	-
+Variability (2nd year)	33.9	0.19	34.7	0.11	26.0	0.11
+Pattern (two years)	32.6	0.53	32.2	1.00	-	-
+Pattern (2nd year)	34.2	0.16	33.0	0.37	24.6	0.27
+Inter. (M&V) (two years)	-	-	36.0	0.28	-	-
+Inter. (M&V) (2nd year)	38.6	0.09	36.9	0.20	26.8	0.18

**Table 7.7.** Assessing the effect of cyclosporin measures after allowing for baseline influences

## 7.6 Discussion

This chapter has described a collaborative project undertaken with clinicians at the Renal Unit at Glasgow's Western Infirmary. The work comprises part of an ongoing investigation into the effects of cyclosporin. There is still considerable debate about which treatment maximises the positive short term effects while guarding against side effects which may affect long term outcome. This decision is not only clouded by the complex nature of the data available but also by other pressures, not the least of which are the interests of the manufacturer of cyclosporin.

This current analysis has used standard survival data analysis but has employed nonparametric modelling to define covariates of interest. Given the nature of

the cyclosporin blood levels data it is hard to imagine how these data could be summarised without the use of some nonparametric modelling. Once again the flexibility and interpretability of smoothing methods have been shown to be well suited to capturing the essence of the relationships between variables.

The results showed that when the cyclosporin variables were considered individually the *variability* in the data values was statistically significant in a proportional hazards model. Furthermore, the *pattern in the doses in the second year* as captured in the monotonicity variables was also found to be significant. The latter result is at least partly a reflection of the clinical response to a deteriorating condition, since an increase in dosage may be prompted by a worsening condition. The former result is harder to interpret, especially since the direction of the effect differs depending on the cyclosporin measure considered. For doses the effect is negative, i.e. increasing the variability measure decreases the hazard function, and vice versa for the levels variability measure. Recall, however, that these two measures are quite different in how they are calculated and what they represent.

When the cyclosporin variables are assessed in the presence of other confounding variables, however, there is no longer any evidence of association between these and the long term outcome. Given the very strong effects of variables such as the *transplant number* and the *age of the donor* these results are perhaps not surprising.

So what are the clinical conclusions to be drawn from such an analysis? To a large degree these results have confirmed pre-existing beliefs regarding the role of cyclosporin dosage. One example is the lack of a significant cyclosporin *magnitude* effect. This non-significant result works both ways in that it suggests there are neither significant positive nor significant negative effects associated with the dosage level prescribed throughout the two years post-transplantation.

Of course, the analysis presented here is clearly not a definitive answer to the 'cyclosporin question' nor was it intended to be. The randomised controlled trials described in the introduction are far better placed to indicate the answer. This study is really a response to the lack of adoption of the recommendations of these randomised studies, since there is a reluctance to remove patients from cyclosporin despite the favourable results of such actions published in previous

studies (MacPhee *et al.* 1998). Given that cyclosporin treatment remains the norm, this analysis has examined if there is evidence of differences in the outcome depending on the degree of exposure to cyclosporin.

The considerable limitation imposed by a non-randomised retrospective study can not be ignored, however. The data itself is limiting in that it is open to potential unknowable influences. Still, it is hoped that this analysis will contribute some insight to the clinical view of post-transplantation treatment as well as illustrate further uses of a nonparametric modelling approach to data analysis.

# Chapter 8

## Conclusions

### 8.1 Introduction

This thesis has considered the task of using fitted nonparametric models to detect and describe the relationships between response and predictor variables in a regression setting. It has built on results from over ten years of focused research into methods of estimation for nonparametric models. Comparisons of model fits have been its focus and methods for these have been proposed in a number of contexts. The aim was to contribute to the understanding of the issues of inference amongst nonparametric models.

Although each chapter has contained within it discussions of the issues raised and the conclusions drawn, it is helpful to summarise the key findings and reflect on the potential for further work. The remaining sections consider three themes of the thesis, respectively: error variance estimation for nonparametric regression models; methods of model comparisons amongst a bivariate class; and extensions and generalisations beyond these special cases. Within each of these areas a brief summary of the main findings is given, followed by a discussion of some of the issues raised and the potential for further work in the area.



## 8.2 Estimation of Error Variance

### 8.2.1 Summary and discussion

Consideration of estimators of  $\sigma^2$  in the nonparametric regression context was prompted by their fundamental role in methods of inference. Although a number of approaches were available in the univariate setting, extensions to the bivariate setting were rare. A strategy of ‘borrowing’ approaches from the parametric regression settings was noted, however the challenges of the nonparametric context deemed it necessary to consider this task afresh with the goal of developing estimators suitable for use in a comparison of bivariate nonparametric model fits.

Two general approaches were adapted and investigated in the bivariate setting. The first were the familiar estimators of  $\sigma^2$  based on RSS of a suitable model fit. The consequences of the unavoidable presence of bias in nonparametric fits was noted and methods were proposed to reduce this effect, namely undersmoothing and double-smoothing. Both of these methods required the specification of a smoothing parameter directly which prompted the search for an estimator of  $\sigma^2$  with low bias which didn’t require an explicit fit to the data. Difference based methods were extended beyond the univariate to the bivariate setting. The estimates were based on interpolating planes using Delaunay triangulations to partition the design space to identify neighbouring points.

The different estimators were investigated over both regular and random designs via simulation studies. These showed that both the undersmoothed RSS based and difference based estimators had small finite sample bias properties. It was observed that the more complicated approaches of the double-smoothed RSS estimator and the ‘weighted triplets’ difference based estimator yielded marginally smaller bias results. Misspecifying the model in the RSS based estimator was shown to increase the bias marginally. The main conclusion was that the difference based estimator offered a viable alternative to the RSS based estimator with the advantage that it didn’t require an explicit model fit and therefore was free of a smoothing parameter.

### 8.2.2 Future work

In terms of further developments, there is scope to consider different configurations and weightings of local points in the difference based estimator. Some work has been done for grid designs (Hall *et al.* 1991) but this idea of ‘optimal configurations and weights’ have not been extended to random designs. There are obvious challenges to accommodate the limitless potential configurations, however, and it may be that the intuitive definition and scaling of neighbouring points given here is close to optimal. Further work is needed to ascertain this.

A natural extension of this work would be to develop difference based estimators for settings beyond two covariates. It is clear, however, from the issues concerning the definition of neighbouring points in the bivariate context, how complicated the methodology becomes in higher dimensions. This is the ultimate advantage of RSS based estimators since they effortlessly and intuitively yield statistics which can be used to estimate  $\sigma^2$  in the context of multiple predictors.

Within the RSS based approach there is room for exploring different approaches to model fitting for the express purpose of  $\sigma^2$  estimation. Some of the potential improvements flagged in Chapter 2 were the use of cross-validation and different types of smoothers to define the fitted surface. Indeed, since there is still much interest in the estimation of nonparametric surfaces it would be beneficial if new (and existing) methods could keep *both* of the unknown ‘quantities’ in view, that is the regression surface *and* the error variance. This would highlight the necessity that both estimation and inference develop in order for nonparametric regression models to achieve their full potential.

## 8.3 Inference Amongst Bivariate Models

### 8.3.1 Summary and discussion

The development of hypothesis tests for comparing models belonging to a class of bivariate nonparametric regression models is at the heart of this thesis. A review of recent results on the asymptotic bias properties of models fitted by local linear regression lead to the suggestion of a direct comparison of the fitted

values (CFV) of competing models as an alternative to the standard comparisons of residual sums of squares. In addition to these comparison of the model fit statistics, the role of the estimator of  $\sigma^2$  and the reference distribution used to assess the observed values of the test statistic were also investigated in a range of settings.

The principle findings of the simulation studies were that there was much to recommend the use of CFV statistics. In the regular grid setting the equivalent asymptotic biases of different model fits were realised in the finite sample properties and therefore the CFV statistic circumvented the bias problems by permitting them to cancel in the comparisons of the model fits. A consequence of this property was that a low-bias estimator of  $\sigma^2$  was of supreme importance. It is in this setting that difference based estimators of  $\sigma^2$  are strongly recommended since they are independent of the model fits (and thus of smoothing parameters) and ensure small bias comparable to that of the best RSS based estimators. When a CFV model comparison statistic and a difference based estimator of  $\sigma^2$  are combined a corrected F reference distribution captures the distribution of the test statistic well.

In the random design setting a different optimal approach was observed. Because the finite sample biases of competing model fits do not cancel exactly in the context of random designs, a minimum bias estimator of  $\sigma^2$  did not return the most consistent results. Rather, the combination of a CFV statistic with an RSS based estimator of  $\sigma^2$  using the full model fit used in the definition of CFV (i.e. no undersmoothing) was found to return consistent results when referred to a corrected F distribution. This was true over a range of smoothing parameters and different settings. The exceptions to this were comparisons between bivariate additive and bivariate smooth models. Here the most consistent results were once again using a CFV statistic and a corrected F distribution but restricted to moderate smoothing parameter levels and using either difference based or moderately undersmoothed RSS based estimators of  $\sigma^2$ .

An underlying assumption of all of the simulations performed was that model comparisons should be made between fits built on centred smoothers and that corresponding smoothing parameters in the two fits should be the same. Although

these approaches were initially motivated by the asymptotic bias results, simulations exploring this underlying philosophy showed that not following these guidelines could result in marked deteriorations in the finite sample behaviour of the test methods. Recently Opsomer and Ruppert (1999) commented on an ‘often overlooked fact’ in the fitting of semiparametric models which required the use of centred smoothers. Since the role of centred smoothers in the task of estimation is ‘overlooked’, it is felt necessary to stress its importance here, together with that of equivalent smoothing parameters, in the context of inference via model comparisons.

Although not a major focus of this current work, the bootstrap approach to model inference in this setting is worth a comment. Several times when this work has been presented in talks and papers it has been suggested that bootstrap methods can be used rather than develop modifications of ‘classical’ approaches to model inference. As one referee put it ‘.. how important is it to derive test statistics that can be compared to prespecified (e.g.  $F$ ) distributions? Why not just simulate the tail probability for the test for a given data set?’. This suggestion, however, overlooks the difficulties associated with deriving an appropriate test statistic. As was noted in Chapter 1, the presence of bias in the nonparametric fits complicates matters, even when bootstrap procedures are used.

This current work’s focus has been the derivation of test statistics which address the problem of bias when nonparametric models are used to assess covariate effects. When using bootstrap techniques the same challenges arise since the fitted values and residuals are distorted by bias. Davison and Hinkley (1997) describe a strategy which involves both undersmoothing and oversmoothing the original fit, but as they note ‘asymptotic theory shows that something along these lines is necessary to make resampling work, but there is no clear guidance as to precise relative values for the tuning constants [degrees of over/under smoothing]’.

The point to stress is that bootstrap approaches **do not** solve the problems bias cause with methods of inference based on nonparametric models. Clearly, however, the power of resampling should not be ignored in future developments in this area but they will almost certainly complement, rather than compete with, methods such as those developed here. For example, the bootstrap methods used with the reef data in Section 4.4 confirmed the use of a corrected  $F$  reference

distribution.

### 8.3.2 Future work

There remain many options available in the task of comparing nonparametric model fits. The following sections describe several ways in which the results could be extended in future work.

#### Extending the scope of the settings investigated

As with any simulation study, only a relatively small number of settings and conditions could be investigated in this current work. Future work in this area primarily lies in extending the simulations to investigate the performance of these tests under a wider range of conditions. A major consideration is the performance of the tests with different forms of underlying regression functions. Varying types and degrees of departures from simple linear models could be investigated, although any investigation of this type is over-shadowed by the limitless forms of the regression function which are possible.

It would also be useful for further simulations to explore different types of dependency amongst the covariates. The properties of nonparametric model fits in the presence of dependent covariates have not been studied in great depth and therefore further developments in inference would need to parallel a greater understanding of the consequences of dependent covariates on the estimation of nonparametric models.

#### Local fits of polynomials with degree $> 1$

Based on the asymptotic bias results which are available, the same methods for model comparisons described here could be used amongst fits produced by local polynomial smoothing with polynomials of degree greater than one. These would have the advantage that the fits are asymptotically unbiased when the underlying true regression function is a polynomial of degree equal to that of the polynomial fitted locally. Furthermore, higher order local fits result in a smaller order of asymptotic bias (Fan, 1992), which is potentially to the advantage of methods of inference.

The flip side, however, is that in general higher order local polynomial fits contain extra variability. The exception to this is an even  $(2q)$  degree polynomial fit compared to the next odd  $(2q + 1)$  degree polynomial fit, in which case there is no difference in asymptotic variance (Ruppert and Wand 1994). On these grounds, local cubic regression would be the next order polynomial fit (beyond local linear fits) to use. However, the asymptotic theory of local cubic fits is not as well developed as that for the local linear approach and thus further work on estimation using higher order polynomials is necessary to facilitate methods of inference.

These guidelines, however, are given on the grounds of optimising the fits from the perspective of estimation not inference. There may therefore be benefits from exploring the use of higher order local polynomials for the purposes of model comparisons. Furthermore, authors such as Cleveland and Loader (1996) make a strong defence of even degree polynomials on the grounds of finite sample properties and therefore their use should not be entirely excluded.

### **Adapting the methods for use with other smoothers**

Note that the methods developed for comparing model fits are all in terms of the smoothing matrices used to generate the fitted models. This suggests that the test can potentially be employed for use with any linear smoother. Given that prejudices do exist in the smoothing community there may be interest in such extensions. However, the performance of the methods was grounded in the asymptotic and finite sample properties of the smooths and therefore these would need to be checked and the methods adapted to alternate smoothers.

It should also be noted, however, that considerable computational effort and storage is required to generate the  $n \times n$  matrices which drive the tests. Furthermore, the smoothing matrices are not always directly available, since often more computationally efficient methods used to fit the models (see Chambers and Hastie (1993) for a detailed description of how nonparametric models are fit in S). In the case of smoothing splines (an obvious alternative to local polynomial fits) this is particularly so since the method is based on an optimisation problem and not local averaging. Therefore, these approaches may not be suited for use

with such smoothers. This may be an argument in favour of local polynomials over splines since to our knowledge no comparable methods have been proposed for comparing smoothing spline model fits.

The methods considered in this current work are also based on explicit solutions to the backfitting algorithm. Recently, estimators based on ‘marginal integration’ have been proposed, e.g. by Linton and Nielsen (1995) and Fan *et al.* (1998). These estimators can be calculated explicitly, unlike backfitting which only yields explicit fits in low dimensional cases. A comparison of backfitting with marginal integration is necessary from purely an estimation perspective but this may lead to further developments in the area of inference. Backfitting, however, is currently by far the most widely adopted method of estimating additive models and therefore this current work is of immediate relevance.

## 8.4 Inference with Broader Classes of Models

### 8.4.1 Summary and discussion

Although the methods of inference were developed in the context of a bivariate class of nonparametric models, extensions beyond these models were considered in Chapter 5. There it was shown that the presence of additional linear terms did not pose a major obstacle to the application of the tests to compare models. Indeed, although the scope of the simulations was scaled down in this context, a consequence of the larger sample size required for models of higher dimension, similar behaviours to those observed in the bivariate case were observed. Over regular grids, a test using the CFV comparison with an undersmoothed RSS based estimator of  $\sigma^2$  and a corrected F reference distribution was found to balance optimally. Similarly, this approach was the best over random designs with the important difference that undersmoothing was not of benefit to the test’s performance.

An important consideration when more covariates are admitted is the relative size of the boundary region of the design points. As the dimensions increase the proportion of points lying near the boundary increases. This has an impact on the properties of the model fits and therefore on methods of inference based on

them. For instance, asymptotic biases at interior and boundary points differ. The methods developed in Chapters 3-5 were motivated largely by the properties of interior points. This reflects the implicit assumption that the majority of points lie in the interior of the design. If the relative size of this region was to decrease (as the dimension of the design space increases) then more attention should focus on the boundary properties which could lead to amendments to the comparison methodology.

The methods of inference developed herein require that certain assumptions hold, e.g. approximate normality of errors and constant variance, in order to yield meaningful results. This reminds us that diagnostic checks need to be made in the course of any analysis based on fitted models. Graphical methods have established themselves as the main approach to model diagnostics and should be used with the methods of model comparisons developed here. As more covariates are admitted to the model space, however, the task of visualising the data becomes increasingly complex. Cleveland (1993) presents an excellent case for the use of visualisation tools in the exploration of the relationships between variables in a regression setting. He discusses the challenges of graphically exploring the relationships amongst four or more variables (so called *hypervariate data*). These challenges, however, must be faced if these methods of inference are to be employed in practice.

#### 8.4.2 Future work

When considering areas of future work in methods of inference amongst extended classes of nonparametric models each of the topics listed in Section 8.3.2 apply. Consideration of the broader class of models, however, does suggest other areas where further developments are necessary.

##### **Incorporating binning techniques**

Issues related to the size of the data sets and the computational resources required to process this much data become increasingly relevant as the number of covariates increase. Binning was mentioned as a way to tackle this problem since it has been successfully incorporated into estimation techniques. Extending binning to



methods of inference is clearly an area of future work. These are currently under development in the univariate context as part of the 'sm' library S-Plus code described in Bowman and Azzalini (1997).

### **Extensions beyond two nonparametric components**

In the research literature on the estimation of semiparametric and additive models the restriction to two covariates is rarely made. Instead models are usually defined as having any number of nonparametric components, although properties are only listed in detail for the special cases of one or two nonparametric terms (see, for instance, Opsomer and Ruppert 1997 and 1999). This partly accounts for why the current work on inference has been restricted to two nonparametric terms. However, developing model inference methods for use with an unlimited number of nonparametric terms is highly desirable. Indeed, there may be scope for this by considering the recursive definitions which define fits in higher (nonparametric) dimensions, although such definitions potentially obscure the properties of the fits noted and used in the course of this work.

### **Generalizations in the GLM sense**

An exciting class of models which is emerging more and more in the applied setting are nonparametric equivalents to Generalized Linear Models (GLMs), namely the Generalised Additive Model (GAM) class introduced comprehensively in Hastie and Tibshirani (1990). These extend additive nonparametric models by introducing a link function other than the identity between the response and the predictors and, most importantly, admit responses with any distribution from the exponential family, i.e. not just normal errors.

The methods of inference proposed for use with GAMs are by analogy with analysis of deviance in the GLM setting. There is, therefore, an obvious need to develop methods of inference which expressly take into account the bias which is present in the smooth versions of covariate effects. This setting does, however, have the extra complication that iterative estimation is inevitable and therefore the properties of estimators are more difficult to derive. This challenge may indeed be the ultimate hindrance to the widespread adoption of GAMs and is

therefore worthy of future research activity.

# Appendix A

## Software

Listed below is a summary of the S-Plus functions created in the course of this work. They build on the functions available in the ‘sm’ library of Bowman and Azzalini (1997). Each function is listed together with its principal arguments and a brief description.

`simdata(s.size,model,grid,var, ...)` This function generates `s.size` data points in the plane, either as a regularly spaced rectangular design of points or one of a number of random designs as specified by the `grid` parameter. Over these design points one of the regression functions listed in Section 2.7.2 is generated according to the `model` parameter. As well as this regression surface, the function also returns simulated responses, i.e. these values with random  $N(0,var)$  errors added.

`sm.matrices(data, model, h1, h2=NA, ...)` This function takes bivariate design points contained in `data` and generates a smoothing matrix corresponding to a nonparametric model from the bivariate class defined in Section 3.2. The smoothing parameter(s) used in the fit are defined by `h1` and `h2`.

`sp.sm.matrices(data, model, ncovar, h, ...)` This function calculates the smoothing matrix which returns the semiparametric additive model fit specified by `model` using the `ncovar` covariates’ values in `data` and smoothing parameter `h` for each of the nonparametric components. NB. The covariates in `data` must be ordered such that the ‘nonparametric’ covariates are listed first.

- `pvalcomp1(data, S.0, S.1, var, ...)` This function compares the model fits defined by smoothing matrices `S.0` and `S.1` applied to the responses in `data`. It returns p-values for the six tests corresponding to the combinations of two comparisons of model fit statistics (CFV and RSSD) and three reference distributions (QF, F, F.cor) using the true value of the error variance, `var`.
- `cmf.matrices(S.0, S.1, ...)` This function takes the smoothing matrices `S.0` and `S.1` which define competing fits and returns a list of two matrices which define the comparison of model fits statistics CFV and RSSD.
- `sigma.matrix(data,model,h1,h2,h3, ...)` This function returns a list of matrices which define estimators of the error variance around the regression surface underlying data. The list consists of difference based estimators and RSS based estimators from the fitted bivariate model using various degrees of smoothing (`h1`, `h2`, `h3`).
- `pvalcomp2(data, A, j, B, k, ...)` This function returns three p-values for comparison of model fit tests (applied to `data`) corresponding to three reference distributions (QF, F, F.cor). The comparison of model fits statistic used is defined by the `j`th matrix contained in `A`, corresponding to either CFV or RSSD. The matrices held in `B` define different estimators of  $\sigma^2$  (Difference of RSS based) with `k` identifying which of these to use.

## Appendix B

### Simulation Results Summarised in Chapter 4

**Table B.1.** Results of simulation study to investigate the size of tests of model comparisons using RSSD and CFV when  $\sigma^2$  is known. Design points form a regular square grid. The results listed are the proportion of 1000 simulated p-values less than 0.05 under the reduced mode.

red.model	full.model	n	sig.sq	h.num	rss.qf	rss.chi	rss.chi.corr	cfv.qf	cfv.chi	cfv.chi.corr
st.line	semi.par.x2	49	0.01	0.15	0.049	0.040	0.051	0.055	0.045	0.057
st.line	semi.par.x2	49	0.01	0.25	0.058	0.044	0.058	0.051	0.041	0.054
st.line	semi.par.x2	49	0.04	0.15	0.050	0.037	0.050	0.047	0.034	0.047
st.line	semi.par.x2	49	0.04	0.25	0.051	0.036	0.052	0.048	0.036	0.048
st.line	semi.par.x2	100	0.01	0.15	0.051	0.046	0.052	0.050	0.043	0.052
st.line	semi.par.x2	100	0.01	0.25	0.057	0.044	0.058	0.050	0.043	0.050
st.line	semi.par.x2	100	0.04	0.15	0.044	0.036	0.045	0.045	0.032	0.046
st.line	semi.par.x2	100	0.04	0.25	0.044	0.030	0.045	0.049	0.037	0.050
1d.sm	2d.am	49	0.01	0.15	0.049	0.042	0.051	0.053	0.040	0.056
1d.sm	2d.am	49	0.01	0.25	0.050	0.038	0.050	0.054	0.041	0.054
1d.sm	2d.am	49	0.04	0.15	0.047	0.038	0.048	0.049	0.035	0.050
1d.sm	2d.am	49	0.04	0.25	0.052	0.040	0.052	0.050	0.043	0.050
1d.sm	2d.am	100	0.01	0.15	0.054	0.041	0.055	0.057	0.042	0.057
1d.sm	2d.am	100	0.01	0.25	0.061	0.047	0.062	0.055	0.043	0.056
1d.sm	2d.am	100	0.04	0.15	0.047	0.036	0.047	0.050	0.039	0.051
1d.sm	2d.am	100	0.04	0.25	0.049	0.041	0.049	0.054	0.039	0.056
1d.sm	semi.par.x1	49	0.01	0.15	0.063	0.063	0.063	0.063	0.063	0.063
1d.sm	semi.par.x1	49	0.01	0.25	0.063	0.063	0.063	0.063	0.063	0.063
1d.sm	semi.par.x1	49	0.04	0.15	0.056	0.056	0.056	0.056	0.056	0.056
1d.sm	semi.par.x1	49	0.04	0.25	0.056	0.056	0.056	0.056	0.056	0.056
1d.sm	semi.par.x1	100	0.01	0.15	0.053	0.053	0.053	0.053	0.053	0.053
1d.sm	semi.par.x1	100	0.01	0.25	0.053	0.053	0.053	0.053	0.053	0.053
1d.sm	semi.par.x1	100	0.04	0.15	0.040	0.040	0.040	0.040	0.040	0.040
1d.sm	semi.par.x1	100	0.04	0.25	0.040	0.040	0.040	0.040	0.040	0.040
semi.par.x1	2d.am	49	0.01	0.15	0.045	0.031	0.047	0.050	0.027	0.050
semi.par.x1	2d.am	49	0.01	0.25	0.045	0.031	0.046	0.050	0.022	0.050
semi.par.x1	2d.am	49	0.04	0.15	0.034	0.027	0.035	0.041	0.020	0.041
semi.par.x1	2d.am	49	0.04	0.25	0.040	0.028	0.042	0.048	0.017	0.048
semi.par.x1	2d.am	100	0.01	0.15	0.056	0.035	0.056	0.056	0.034	0.057
semi.par.x1	2d.am	100	0.01	0.25	0.059	0.040	0.060	0.065	0.024	0.065
semi.par.x1	2d.am	100	0.04	0.15	0.039	0.024	0.042	0.046	0.026	0.046
semi.par.x1	2d.am	100	0.04	0.25	0.041	0.026	0.041	0.045	0.018	0.046
2d.am	2d.sm	49	0.01	0.15	0.059	0.028	0.059	0.037	0.020	0.040
2d.am	2d.sm	49	0.01	0.25	0.041	0.020	0.045	0.037	0.016	0.040
2d.am	2d.sm	49	0.04	0.15	0.039	0.022	0.040	0.042	0.022	0.045
2d.am	2d.sm	49	0.04	0.25	0.042	0.020	0.047	0.033	0.021	0.033
2d.am	2d.sm	100	0.01	0.15	0.054	0.029	0.055	0.051	0.024	0.056
2d.am	2d.sm	100	0.01	0.25	0.056	0.026	0.058	0.051	0.020	0.051
2d.am	2d.sm	100	0.04	0.15	0.056	0.024	0.061	0.055	0.027	0.058
2d.am	2d.sm	100	0.04	0.25	0.051	0.026	0.055	0.056	0.024	0.057

**Table B.2.** Results of simulation study to investigate the power of tests for model comparisons when  $\sigma^2$  is known. Design points form a regular square grid. The results listed are the proportion of 500 simulated p-values less than 0.05 under the full model.

	red.model	full.model	n	sig.sq	h.num	rss.qf	rss.chi	rss.chi.corr	cfv.qf	cfv.chi	cfv.chi.corr
1	st.line	semi.par.x2	49	0.01	0.15	0.924	0.906	0.926	0.918	0.902	0.920
2	st.line	semi.par.x2	49	0.01	0.25	0.922	0.904	0.924	0.918	0.910	0.918
3	st.line	semi.par.x2	49	0.04	0.15	0.326	0.292	0.326	0.352	0.294	0.354
4	st.line	semi.par.x2	49	0.04	0.25	0.352	0.322	0.352	0.384	0.360	0.390
5	st.line	semi.par.x2	100	0.01	0.15	1.000	0.998	1.000	1.000	0.996	1.000
6	st.line	semi.par.x2	100	0.01	0.25	1.000	1.000	1.000	0.998	0.998	0.998
7	st.line	semi.par.x2	100	0.04	0.15	0.580	0.542	0.584	0.616	0.554	0.620
8	st.line	semi.par.x2	100	0.04	0.25	0.626	0.572	0.632	0.656	0.608	0.658
9	1d.sm	2d.am	49	0.01	0.15	0.720	0.700	0.730	0.750	0.714	0.752
10	1d.sm	2d.am	49	0.01	0.25	0.752	0.722	0.752	0.772	0.740	0.774
11	1d.sm	2d.am	49	0.04	0.15	0.262	0.226	0.264	0.250	0.210	0.252
12	1d.sm	2d.am	49	0.04	0.25	0.252	0.220	0.252	0.270	0.240	0.270
13	1d.sm	2d.am	100	0.01	0.15	0.990	0.976	0.990	0.984	0.980	0.986
14	1d.sm	2d.am	100	0.01	0.25	0.986	0.980	0.986	0.986	0.976	0.986
15	1d.sm	2d.am	100	0.04	0.15	0.424	0.384	0.426	0.422	0.382	0.426
16	1d.sm	2d.am	100	0.04	0.25	0.424	0.392	0.428	0.460	0.414	0.462
17	1d.sm	semi.par.x1	49	0.01	0.15	0.966	0.966	0.966	0.966	0.966	0.966
18	1d.sm	semi.par.x1	49	0.01	0.25	0.966	0.966	0.966	0.966	0.966	0.966
19	1d.sm	semi.par.x1	49	0.04	0.15	0.476	0.476	0.476	0.476	0.476	0.476
20	1d.sm	semi.par.x1	49	0.04	0.25	0.476	0.476	0.476	0.476	0.476	0.476
21	1d.sm	semi.par.x1	100	0.01	0.15	0.998	0.998	0.998	0.998	0.998	0.998
22	1d.sm	semi.par.x1	100	0.01	0.25	0.998	0.998	0.998	0.998	0.998	0.998
23	1d.sm	semi.par.x1	100	0.04	0.15	0.734	0.734	0.734	0.734	0.734	0.734
24	1d.sm	semi.par.x1	100	0.04	0.25	0.734	0.734	0.734	0.734	0.734	0.734
25	semi.par.x1	2d.am	49	0.01	0.15	0.336	0.278	0.338	0.260	0.192	0.260
26	semi.par.x1	2d.am	49	0.01	0.25	0.308	0.238	0.308	0.314	0.184	0.314
27	semi.par.x1	2d.am	49	0.04	0.15	0.110	0.078	0.112	0.084	0.052	0.084
28	semi.par.x1	2d.am	49	0.04	0.25	0.096	0.072	0.096	0.096	0.056	0.096
29	semi.par.x1	2d.am	100	0.01	0.15	0.558	0.512	0.564	0.468	0.372	0.474
30	semi.par.x1	2d.am	100	0.01	0.25	0.532	0.458	0.534	0.542	0.382	0.542
31	semi.par.x1	2d.am	100	0.04	0.15	0.168	0.126	0.174	0.150	0.104	0.154
32	semi.par.x1	2d.am	100	0.04	0.25	0.168	0.136	0.168	0.174	0.102	0.176
33	2d.am	2d.sm	49	0.01	0.15	0.628	0.524	0.636	0.438	0.338	0.446
34	2d.am	2d.sm	49	0.01	0.25	0.370	0.282	0.378	0.262	0.154	0.266
35	2d.am	2d.sm	49	0.04	0.15	0.180	0.110	0.182	0.106	0.070	0.120
36	2d.am	2d.sm	49	0.04	0.25	0.116	0.070	0.118	0.080	0.054	0.082
37	2d.am	2d.sm	100	0.01	0.15	0.946	0.916	0.946	0.874	0.762	0.880
38	2d.am	2d.sm	100	0.01	0.25	0.744	0.636	0.754	0.492	0.316	0.510
39	2d.am	2d.sm	100	0.04	0.15	0.246	0.158	0.260	0.188	0.116	0.196
40	2d.am	2d.sm	100	0.04	0.25	0.170	0.094	0.174	0.116	0.058	0.120

**Table B.3.** Definition of ‘Setups’ summarised in Tables B.4 - B.6

Setup	Reduced Model	Full Model	n	$\sigma^2$	h
1	st.line	semi.par.x2	100	0.01	0.1
2	st.line	semi.par.x2	100	0.01	0.2
3	st.line	semi.par.x2	100	0.0025	0.1
4	st.line	semi.par.x2	100	0.0025	0.2
5	st.line	semi.par.x2	49	0.01	0.15
6	st.line	semi.par.x2	49	0.01	0.25
7	st.line	semi.par.x2	49	0.0025	0.15
8	st.line	semi.par.x2	49	0.0025	0.25
9	1d.sm	2d.am	100	0.01	0.1
10	1d.sm	2d.am	100	0.01	0.2
11	1d.sm	2d.am	100	0.0025	0.1
12	1d.sm	2d.am	100	0.0025	0.2
13	1d.sm	2d.am	49	0.01	0.15
14	1d.sm	2d.am	49	0.01	0.25
15	1d.sm	2d.am	49	0.0025	0.15
16	1d.sm	2d.am	49	0.0025	0.25
17	1d.sm	semi.par.x1	100	0.01	0.1
18	1d.sm	semi.par.x1	100	0.01	0.2
19	1d.sm	semi.par.x1	100	0.0025	0.1
20	1d.sm	semi.par.x1	100	0.0025	0.2
21	1d.sm	semi.par.x1	49	0.01	0.15
22	1d.sm	semi.par.x1	49	0.01	0.25
23	1d.sm	semi.par.x1	49	0.0025	0.15
24	1d.sm	semi.par.x1	49	0.0025	0.25
25	semi.par.x1	2d.am	100	0.01	0.1
26	semi.par.x1	2d.am	100	0.01	0.2
27	semi.par.x1	2d.am	100	0.0025	0.1
28	semi.par.x1	2d.am	100	0.0025	0.2
29	semi.par.x1	2d.am	49	0.01	0.15
30	semi.par.x1	2d.am	49	0.01	0.25
31	semi.par.x1	2d.am	49	0.0025	0.15
32	semi.par.x1	2d.am	49	0.0025	0.25
33	2d.am	2d.sm	100	0.01	0.1
34	2d.am	2d.sm	100	0.01	0.2
35	2d.am	2d.sm	100	0.0025	0.1
36	2d.am	2d.sm	100	0.0025	0.2
37	2d.am	2d.sm	49	0.01	0.15
38	2d.am	2d.sm	49	0.01	0.25
39	2d.am	2d.sm	49	0.0025	0.15
40	2d.am	2d.sm	49	0.0025	0.25



**Table B.4.** Empirical sizes of 8 model comparison tests, each using the corrected F reference distribution. 500 simulated data sets over regular grids were used, generated under the conditions listed in Table B.3

Setup	CFV				RSSD			
	Diff	h.num/4	h.num/2	h.num	Diff	h.num/4	h.num/2	h.num
1	0.062	0.054	0.054	0.056	0.052	0.056	0.052	0.048
2	0.060	0.056	0.054	0.054	0.052	0.050	0.052	0.050
3	0.044	0.034	0.034	0.034	0.050	0.048	0.046	0.042
4	0.032	0.034	0.034	0.032	0.048	0.044	0.046	0.042
5	0.036	0.044	0.046	0.046	0.048	0.052	0.050	0.042
6	0.028	0.046	0.044	0.044	0.042	0.048	0.048	0.046
7	0.054	0.056	0.056	0.056	0.046	0.064	0.064	0.062
8	0.052	0.052	0.052	0.048	0.048	0.058	0.066	0.054
9	0.056	0.038	0.028	0.008	0.056	0.048	0.038	0.006
10	0.058	0.036	0.010	0.002	0.050	0.032	0.006	0.000
11	0.060	0.072	0.046	0.000	0.068	0.068	0.046	0.000
12	0.062	0.048	0.000	0.000	0.058	0.046	0.000	0.000
13	0.044	0.054	0.036	0.002	0.048	0.050	0.032	0.000
14	0.054	0.046	0.004	0.002	0.044	0.048	0.004	0.000
15	0.028	0.036	0.010	0.000	0.038	0.044	0.008	0.000
16	0.032	0.022	0.000	0.000	0.040	0.028	0.000	0.000
17	0.044	0.054	0.050	0.014	0.044	0.054	0.050	0.014
18	0.044	0.050	0.014	0.002	0.044	0.050	0.014	0.002
19	0.056	0.062	0.044	0.002	0.056	0.062	0.044	0.002
20	0.056	0.044	0.002	0.000	0.056	0.044	0.002	0.000
21	0.054	0.052	0.044	0.002	0.054	0.052	0.044	0.002
22	0.054	0.050	0.006	0.000	0.054	0.050	0.006	0.000
23	0.038	0.048	0.016	0.000	0.038	0.048	0.016	0.000
24	0.038	0.042	0.000	0.000	0.038	0.042	0.000	0.000
25	0.030	0.036	0.036	0.024	0.036	0.042	0.040	0.020
26	0.034	0.034	0.024	0.012	0.028	0.038	0.026	0.006
27	0.052	0.038	0.034	0.006	0.042	0.040	0.038	0.004
28	0.044	0.030	0.006	0.002	0.048	0.034	0.006	0.000
29	0.062	0.060	0.052	0.020	0.054	0.070	0.056	0.014
30	0.060	0.060	0.034	0.016	0.064	0.058	0.028	0.012
31	0.028	0.034	0.014	0.000	0.034	0.038	0.018	0.000
32	0.040	0.034	0.000	0.000	0.026	0.028	0.000	0.000
33	0.034	0.016	0.010	0.002	0.022	0.002	0.000	0.000
34	0.044	0.030	0.020	0.000	0.028	0.014	0.000	0.000
35	0.046	0.000	0.000	0.000	0.042	0.000	0.000	0.000
36	0.042	0.004	0.000	0.000	0.048	0.000	0.000	0.000
37	0.034	0.006	0.004	0.002	0.028	0.000	0.000	0.000
38	0.034	0.006	0.004	0.000	0.032	0.006	0.002	0.000
39	0.038	0.000	0.000	0.000	0.026	0.000	0.000	0.000
40	0.042	0.000	0.000	0.000	0.044	0.000	0.000	0.000

**Table B.5.** Empirical sizes of 8 model comparison tests, each using the ‘QF’ reference distribution. 500 simulated data sets over regular grids were used, generated under the conditions listed in Table B.3

Setup	CFV				RSSD			
	Diff	h.num/4	h.num/2	h.num	Diff	h.num/4	h.num/2	h.num
1	0.058	0.052	0.052	0.050	0.050	0.048	0.048	0.048
2	0.048	0.056	0.054	0.052	0.048	0.048	0.048	0.050
3	0.042	0.034	0.034	0.034	0.048	0.046	0.046	0.044
4	0.028	0.030	0.028	0.028	0.046	0.044	0.044	0.042
5	0.010	0.040	0.040	0.044	0.008	0.050	0.050	0.048
6	0.018	0.044	0.042	0.042	0.014	0.046	0.046	0.046
7	0.024	0.054	0.054	0.054	0.012	0.062	0.064	0.064
8	0.028	0.048	0.046	0.042	0.022	0.054	0.056	0.054
9	0.056	0.034	0.026	0.008	0.048	0.046	0.038	0.006
10	0.046	0.032	0.010	0.002	0.050	0.032	0.006	0.000
11	0.058	0.068	0.046	0.000	0.066	0.068	0.046	0.000
12	0.056	0.046	0.000	0.000	0.058	0.046	0.000	0.000
13	0.022	0.046	0.036	0.002	0.008	0.050	0.032	0.002
14	0.024	0.042	0.004	0.002	0.016	0.048	0.004	0.000
15	0.006	0.028	0.010	0.000	0.002	0.040	0.008	0.000
16	0.012	0.020	0.000	0.000	0.004	0.028	0.000	0.000
17	0.030	0.050	0.046	0.010	0.030	0.050	0.046	0.010
18	0.030	0.046	0.010	0.002	0.030	0.046	0.010	0.002
19	0.050	0.056	0.042	0.002	0.050	0.056	0.042	0.002
20	0.050	0.042	0.002	0.000	0.050	0.042	0.002	0.000
21	0.042	0.046	0.036	0.002	0.042	0.046	0.036	0.002
22	0.042	0.044	0.006	0.000	0.042	0.044	0.006	0.000
23	0.026	0.040	0.014	0.000	0.026	0.040	0.014	0.000
24	0.026	0.034	0.000	0.000	0.026	0.034	0.000	0.000
25	0.030	0.034	0.034	0.024	0.034	0.036	0.034	0.020
26	0.032	0.034	0.022	0.012	0.022	0.034	0.022	0.006
27	0.048	0.034	0.032	0.006	0.042	0.040	0.038	0.004
28	0.040	0.026	0.006	0.000	0.046	0.032	0.006	0.000
29	0.018	0.056	0.048	0.020	0.022	0.060	0.054	0.016
30	0.040	0.056	0.030	0.016	0.020	0.054	0.026	0.012
31	0.010	0.020	0.014	0.000	0.006	0.034	0.016	0.000
32	0.016	0.030	0.000	0.000	0.010	0.022	0.000	0.000
33	0.014	0.016	0.010	0.002	0.014	0.004	0.000	0.000
34	0.044	0.030	0.016	0.000	0.020	0.014	0.000	0.000
35	0.022	0.000	0.000	0.000	0.014	0.000	0.000	0.000
36	0.038	0.004	0.000	0.000	0.030	0.000	0.000	0.000
37	0.008	0.004	0.004	0.002	0.002	0.000	0.000	0.000
38	0.008	0.006	0.002	0.000	0.006	0.004	0.002	0.000
39	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000
40	0.018	0.000	0.000	0.000	0.004	0.000	0.000	0.000

**Table B.6.** Empirical sizes of 8 model comparison tests, each using the (unadjusted) F reference distribution. 500 simulated data sets over regular grids were used, generated under the conditions listed in Table B.3

Setup	CFV				RSSD			
	Diff	h.num/4	h.num/2	h.num	Diff	h.num/4	h.num/2	h.num
1	0	0.034	0.034	0.032	0	0.042	0.042	0.038
2	0	0.040	0.038	0.038	0	0.034	0.030	0.032
3	0	0.030	0.030	0.028	0	0.032	0.032	0.032
4	0	0.022	0.020	0.020	0	0.032	0.032	0.028
5	0	0.038	0.038	0.036	0	0.040	0.040	0.036
6	0	0.036	0.038	0.038	0	0.042	0.042	0.038
7	0	0.048	0.048	0.046	0	0.052	0.052	0.042
8	0	0.044	0.042	0.038	0	0.050	0.044	0.046
9	0	0.026	0.024	0.008	0	0.036	0.024	0.006
10	0	0.032	0.010	0.000	0	0.024	0.006	0.000
11	0	0.052	0.042	0.000	0	0.058	0.038	0.000
12	0	0.038	0.000	0.000	0	0.040	0.000	0.000
13	0	0.046	0.032	0.002	0	0.044	0.030	0.000
14	0	0.042	0.004	0.002	0	0.048	0.004	0.000
15	0	0.022	0.008	0.000	0	0.036	0.006	0.000
16	0	0.020	0.000	0.000	0	0.024	0.000	0.000
17	0	0.054	0.050	0.014	0	0.054	0.050	0.014
18	0	0.050	0.014	0.002	0	0.050	0.014	0.002
19	0	0.062	0.044	0.002	0	0.062	0.044	0.002
20	0	0.044	0.002	0.000	0	0.044	0.002	0.000
21	0	0.052	0.044	0.002	0	0.052	0.044	0.002
22	0	0.050	0.006	0.000	0	0.050	0.006	0.000
23	0	0.048	0.016	0.000	0	0.048	0.016	0.000
24	0	0.042	0.000	0.000	0	0.042	0.000	0.000
25	0	0.026	0.026	0.014	0	0.026	0.026	0.012
26	0	0.020	0.012	0.000	0	0.024	0.016	0.000
27	0	0.020	0.018	0.004	0	0.034	0.030	0.004
28	0	0.016	0.004	0.000	0	0.022	0.004	0.000
29	0	0.034	0.032	0.010	0	0.050	0.040	0.010
30	0	0.030	0.014	0.006	0	0.044	0.020	0.008
31	0	0.014	0.008	0.000	0	0.026	0.012	0.000
32	0	0.010	0.000	0.000	0	0.014	0.000	0.000
33	0	0.000	0.000	0.000	0	0.000	0.000	0.000
34	0	0.000	0.004	0.000	0	0.000	0.000	0.000
35	0	0.000	0.000	0.000	0	0.000	0.000	0.000
36	0	0.000	0.000	0.000	0	0.000	0.000	0.000
37	0	0.000	0.000	0.000	0	0.000	0.000	0.000
38	0	0.000	0.000	0.000	0	0.000	0.000	0.000
39	0	0.000	0.000	0.000	0	0.000	0.000	0.000
40	0	0.000	0.000	0.000	0	0.000	0.000	0.000

**Table B.7.** Settings used in the simulation study over random designs using known  $\sigma^2$  (results reported in Section 4.3.1)

	nsim	red.model	full.model	n	sig.sq	h.num
1	500	st.line	semi.par.x2	49	0.01	0.05
2	500	st.line	semi.par.x2	49	0.01	0.15
3	500	st.line	semi.par.x2	49	0.04	0.05
4	500	st.line	semi.par.x2	49	0.04	0.15
5	500	st.line	semi.par.x2	100	0.01	0.05
6	500	st.line	semi.par.x2	100	0.01	0.15
7	500	st.line	semi.par.x2	100	0.04	0.05
8	500	st.line	semi.par.x2	100	0.04	0.15
9	500	1d.sm	2d.am	49	0.01	0.05
10	500	1d.sm	2d.am	49	0.01	0.15
11	500	1d.sm	2d.am	49	0.04	0.05
12	500	1d.sm	2d.am	49	0.04	0.15
13	500	1d.sm	2d.am	100	0.01	0.05
14	500	1d.sm	2d.am	100	0.01	0.15
15	500	1d.sm	2d.am	100	0.04	0.05
16	500	1d.sm	2d.am	100	0.04	0.15
17	500	1d.sm	semi.par.x1	49	0.01	0.05
18	500	1d.sm	semi.par.x1	49	0.01	0.15
19	500	1d.sm	semi.par.x1	49	0.04	0.05
20	500	1d.sm	semi.par.x1	49	0.04	0.15
21	500	1d.sm	semi.par.x1	100	0.01	0.05
22	500	1d.sm	semi.par.x1	100	0.01	0.15
23	500	1d.sm	semi.par.x1	100	0.04	0.05
24	500	1d.sm	semi.par.x1	100	0.04	0.15
25	500	semi.par.x1	2d.am	49	0.01	0.05
26	500	semi.par.x1	2d.am	49	0.01	0.15
27	500	semi.par.x1	2d.am	49	0.04	0.05
28	500	semi.par.x1	2d.am	49	0.04	0.15
29	500	semi.par.x1	2d.am	100	0.01	0.05
30	500	semi.par.x1	2d.am	100	0.01	0.15
31	500	semi.par.x1	2d.am	100	0.04	0.05
32	500	semi.par.x1	2d.am	100	0.04	0.15
33	500	2d.am	2d.sm	49	0.01	0.05
34	500	2d.am	2d.sm	49	0.01	0.15
35	500	2d.am	2d.sm	49	0.04	0.05
36	500	2d.am	2d.sm	49	0.04	0.15
37	500	2d.am	2d.sm	100	0.01	0.05
38	500	2d.am	2d.sm	100	0.01	0.15
39	500	2d.am	2d.sm	100	0.04	0.05
40	500	2d.am	2d.sm	100	0.04	0.15

**Table B.8.** Results of simulation study to investigate the size of tests of model comparisons using RSSD and CFV when  $\sigma^2$  is known. Design points are samples from a uniform distribution over  $[0, 1]^2$ . The results listed are the proportion of 500 simulated p-values less than 0.05.

	red.model	full.model	n	sig.sq	h.num	rss.qf	rss.chi	rss.chi.corr	cfv.qf	cfv.chi	cfv.chi.corr
1	st.line	semi.par.x2	49	0.01	0.05	0.090	0.072	0.090	0.078	0.062	0.078
2	st.line	semi.par.x2	49	0.01	0.15	0.070	0.064	0.070	0.058	0.046	0.060
3	st.line	semi.par.x2	49	0.04	0.05	0.100	0.080	0.100	0.068	0.058	0.068
4	st.line	semi.par.x2	49	0.04	0.15	0.068	0.056	0.068	0.046	0.040	0.048
5	st.line	semi.par.x2	100	0.01	0.05	0.096	0.086	0.096	0.062	0.050	0.064
6	st.line	semi.par.x2	100	0.01	0.15	0.048	0.040	0.048	0.040	0.030	0.040
7	st.line	semi.par.x2	100	0.04	0.05	0.064	0.062	0.064	0.042	0.038	0.044
8	st.line	semi.par.x2	100	0.04	0.15	0.036	0.036	0.036	0.050	0.038	0.050
9	1d.sm	2d.am	49	0.01	0.05	0.130	0.108	0.132	0.114	0.092	0.114
10	1d.sm	2d.am	49	0.01	0.15	0.226	0.196	0.226	0.224	0.194	0.224
11	1d.sm	2d.am	49	0.04	0.05	0.132	0.100	0.134	0.086	0.072	0.086
12	1d.sm	2d.am	49	0.04	0.15	0.088	0.080	0.088	0.066	0.058	0.070
13	1d.sm	2d.am	100	0.01	0.05	0.116	0.102	0.116	0.100	0.080	0.100
14	1d.sm	2d.am	100	0.01	0.15	0.224	0.200	0.228	0.266	0.232	0.268
15	1d.sm	2d.am	100	0.04	0.05	0.098	0.068	0.098	0.092	0.062	0.092
16	1d.sm	2d.am	100	0.04	0.15	0.098	0.076	0.100	0.094	0.080	0.094
17	1d.sm	semi.par.x1	49	0.01	0.05	0.056	0.056	0.056	0.062	0.060	0.062
18	1d.sm	semi.par.x1	49	0.01	0.15	0.122	0.122	0.122	0.164	0.164	0.164
19	1d.sm	semi.par.x1	49	0.04	0.05	0.040	0.040	0.040	0.042	0.038	0.042
20	1d.sm	semi.par.x1	49	0.04	0.15	0.056	0.054	0.056	0.062	0.060	0.062
21	1d.sm	semi.par.x1	100	0.01	0.05	0.062	0.060	0.062	0.060	0.058	0.060
22	1d.sm	semi.par.x1	100	0.01	0.15	0.092	0.092	0.092	0.136	0.136	0.136
23	1d.sm	semi.par.x1	100	0.04	0.05	0.076	0.076	0.076	0.076	0.070	0.076
24	1d.sm	semi.par.x1	100	0.04	0.15	0.078	0.074	0.078	0.084	0.082	0.084
25	semi.par.x1	2d.am	49	0.01	0.05	0.126	0.090	0.126	0.100	0.068	0.100
26	semi.par.x1	2d.am	49	0.01	0.15	0.088	0.066	0.088	0.098	0.056	0.098
27	semi.par.x1	2d.am	49	0.04	0.05	0.134	0.112	0.138	0.114	0.094	0.116
28	semi.par.x1	2d.am	49	0.04	0.15	0.084	0.062	0.084	0.072	0.040	0.074
29	semi.par.x1	2d.am	100	0.01	0.05	0.100	0.082	0.104	0.106	0.060	0.106
30	semi.par.x1	2d.am	100	0.01	0.15	0.106	0.082	0.110	0.100	0.068	0.102
31	semi.par.x1	2d.am	100	0.04	0.05	0.110	0.092	0.112	0.100	0.062	0.104
32	semi.par.x1	2d.am	100	0.04	0.15	0.070	0.056	0.074	0.068	0.050	0.070
33	2d.am	2d.sm	49	0.01	0.05	NA	NA	NA	NA	NA	NA
34	2d.am	2d.sm	49	0.01	0.15	0.238	0.168	0.250	0.148	0.090	0.154
35	2d.am	2d.sm	49	0.04	0.05	NA	NA	NA	NA	NA	NA
36	2d.am	2d.sm	49	0.04	0.15	0.156	0.110	0.156	0.120	0.060	0.128
37	2d.am	2d.sm	100	0.01	0.05	0.264	0.244	0.264	0.240	0.200	0.240
38	2d.am	2d.sm	100	0.01	0.15	0.300	0.236	0.312	0.164	0.108	0.170
39	2d.am	2d.sm	100	0.04	0.05	0.258	0.218	0.262	0.212	0.184	0.216
40	2d.am	2d.sm	100	0.04	0.15	0.142	0.100	0.144	0.080	0.048	0.082

**Table B.9.** Results of simulation study to investigate the size of tests of model comparisons using RSSD and CFV when  $\sigma^2$  is known. Design points are samples from a bivariate normal distribution with  $\mu_1 = \mu_2 = 0.5$  and  $\sigma_1^2 = \sigma_2^2 = 0.15$  and  $\rho = 0$ . The results listed are the proportion of 500 simulated p-values less than 0.05.

	red.model	full.model	n	sig.sq	h.num	rss.qf	rss.chi	rss.chi.corr	cfv.qf	cfv.chi	cfv.chi.corr
1	st.line	semi.par.x2	49	0.01	0.05	0.074	0.068	0.074	0.070	0.064	0.072
2	st.line	semi.par.x2	49	0.01	0.15	0.054	0.050	0.056	0.056	0.048	0.056
3	st.line	semi.par.x2	49	0.04	0.05	0.064	0.058	0.064	0.060	0.046	0.060
4	st.line	semi.par.x2	49	0.04	0.15	0.050	0.042	0.050	0.050	0.038	0.050
5	st.line	semi.par.x2	100	0.01	0.05	0.062	0.056	0.062	0.054	0.048	0.054
6	st.line	semi.par.x2	100	0.01	0.15	0.060	0.052	0.060	0.058	0.046	0.058
7	st.line	semi.par.x2	100	0.04	0.05	0.070	0.058	0.070	0.058	0.048	0.062
8	st.line	semi.par.x2	100	0.04	0.15	0.048	0.034	0.048	0.042	0.036	0.044
9	1d.sm	2d.am	49	0.01	0.05	0.114	0.094	0.114	0.094	0.076	0.096
10	1d.sm	2d.am	49	0.01	0.15	0.166	0.142	0.166	0.192	0.168	0.194
11	1d.sm	2d.am	49	0.04	0.05	0.086	0.070	0.086	0.086	0.060	0.086
12	1d.sm	2d.am	49	0.04	0.15	0.080	0.070	0.080	0.094	0.076	0.096
13	1d.sm	2d.am	100	0.01	0.05	0.094	0.084	0.096	0.094	0.082	0.094
14	1d.sm	2d.am	100	0.01	0.15	0.170	0.130	0.172	0.196	0.166	0.198
15	1d.sm	2d.am	100	0.04	0.05	0.092	0.072	0.094	0.070	0.050	0.074
16	1d.sm	2d.am	100	0.04	0.15	0.080	0.056	0.080	0.070	0.068	0.070
17	1d.sm	semi.par.x1	49	0.01	0.05	0.042	0.040	0.042	0.048	0.048	0.048
18	1d.sm	semi.par.x1	49	0.01	0.15	0.086	0.086	0.086	0.138	0.134	0.138
19	1d.sm	semi.par.x1	49	0.04	0.05	0.050	0.050	0.050	0.052	0.052	0.052
20	1d.sm	semi.par.x1	49	0.04	0.15	0.058	0.058	0.058	0.066	0.064	0.066
21	1d.sm	semi.par.x1	100	0.01	0.05	0.044	0.044	0.044	0.060	0.058	0.060
22	1d.sm	semi.par.x1	100	0.01	0.15	0.086	0.086	0.086	0.156	0.156	0.156
23	1d.sm	semi.par.x1	100	0.04	0.05	0.040	0.040	0.040	0.042	0.042	0.042
24	1d.sm	semi.par.x1	100	0.04	0.15	0.044	0.042	0.044	0.050	0.048	0.050
25	semi.par.x1	2d.am	49	0.01	0.05	0.098	0.084	0.098	0.078	0.054	0.080
26	semi.par.x1	2d.am	49	0.01	0.15	0.072	0.052	0.074	0.086	0.036	0.086
27	semi.par.x1	2d.am	49	0.04	0.05	0.088	0.058	0.088	0.056	0.042	0.060
28	semi.par.x1	2d.am	49	0.04	0.15	0.050	0.036	0.050	0.046	0.030	0.046
29	semi.par.x1	2d.am	100	0.01	0.05	0.080	0.060	0.082	0.064	0.048	0.068
30	semi.par.x1	2d.am	100	0.01	0.15	0.066	0.062	0.068	0.086	0.038	0.088
31	semi.par.x1	2d.am	100	0.04	0.05	0.100	0.080	0.102	0.074	0.046	0.074
32	semi.par.x1	2d.am	100	0.04	0.15	0.056	0.036	0.056	0.056	0.026	0.056
33	2d.am	2d.sm	49	0.01	0.05	0.182	0.144	0.184	0.138	0.090	0.138
34	2d.am	2d.sm	49	0.01	0.15	0.182	0.116	0.190	0.138	0.044	0.140
35	2d.am	2d.sm	49	0.04	0.05	0.162	0.136	0.162	0.140	0.102	0.140
36	2d.am	2d.sm	49	0.04	0.15	0.082	0.050	0.092	0.070	0.024	0.070
37	2d.am	2d.sm	100	0.01	0.05	0.188	0.140	0.188	0.140	0.080	0.144
38	2d.am	2d.sm	100	0.01	0.15	0.226	0.142	0.226	0.160	0.058	0.160
39	2d.am	2d.sm	100	0.04	0.05	0.198	0.160	0.198	0.146	0.088	0.148
40	2d.am	2d.sm	100	0.04	0.15	0.090	0.054	0.092	0.062	0.028	0.066

**Table B.10.** Results of simulation study to investigate the size of tests of model comparisons using RSSD and CFV when  $\sigma^2$  is known. Design points are samples from a bivariate normal distribution with  $\mu_1 = \mu_2 = 0.5$  and  $\sigma_1^2 = \sigma_2^2 = 0.15$  and  $\rho = 0.5$ . The results listed are the proportion of 500 simulated p-values less than 0.05.

	red.model	full.model	n	sig.sq	h.num	rss.qf	rss.chi	rss.chi.corr	cfv.qf	cfv.chi	cfv.chi.corr
1	st.line	semi.par.x2	49	0.01	0.05	0.070	0.058	0.070	0.060	0.052	0.060
2	st.line	semi.par.x2	49	0.01	0.15	0.060	0.048	0.060	0.050	0.040	0.052
3	st.line	semi.par.x2	49	0.04	0.05	0.076	0.062	0.078	0.062	0.046	0.064
4	st.line	semi.par.x2	49	0.04	0.15	0.066	0.054	0.068	0.068	0.058	0.068
5	st.line	semi.par.x2	100	0.01	0.05	0.074	0.066	0.074	0.050	0.046	0.056
6	st.line	semi.par.x2	100	0.01	0.15	0.042	0.032	0.042	0.038	0.032	0.040
7	st.line	semi.par.x2	100	0.04	0.05	0.080	0.070	0.082	0.064	0.052	0.064
8	st.line	semi.par.x2	100	0.04	0.15	0.062	0.040	0.062	0.054	0.046	0.054
9	1d.sm	2d.am	49	0.01	0.05	0.080	0.068	0.080	0.102	0.084	0.104
10	1d.sm	2d.am	49	0.01	0.15	0.144	0.124	0.144	0.220	0.196	0.224
11	1d.sm	2d.am	49	0.04	0.05	0.088	0.072	0.088	0.082	0.068	0.088
12	1d.sm	2d.am	49	0.04	0.15	0.068	0.050	0.070	0.064	0.048	0.064
13	1d.sm	2d.am	100	0.01	0.05	0.088	0.072	0.088	0.104	0.096	0.104
14	1d.sm	2d.am	100	0.01	0.15	0.174	0.142	0.182	0.268	0.222	0.268
15	1d.sm	2d.am	100	0.04	0.05	0.080	0.066	0.080	0.082	0.064	0.082
16	1d.sm	2d.am	100	0.04	0.15	0.076	0.062	0.076	0.074	0.070	0.074
17	1d.sm	semi.par.x1	49	0.01	0.05	0.058	0.058	0.058	0.074	0.068	0.074
18	1d.sm	semi.par.x1	49	0.01	0.15	0.096	0.094	0.096	0.144	0.144	0.144
19	1d.sm	semi.par.x1	49	0.04	0.05	0.068	0.068	0.068	0.064	0.064	0.064
20	1d.sm	semi.par.x1	49	0.04	0.15	0.058	0.058	0.058	0.082	0.078	0.082
21	1d.sm	semi.par.x1	100	0.01	0.05	0.054	0.054	0.054	0.060	0.058	0.060
22	1d.sm	semi.par.x1	100	0.01	0.15	0.104	0.104	0.104	0.168	0.166	0.168
23	1d.sm	semi.par.x1	100	0.04	0.05	0.030	0.030	0.030	0.034	0.030	0.034
24	1d.sm	semi.par.x1	100	0.04	0.15	0.034	0.034	0.034	0.042	0.042	0.042
25	semi.par.x1	2d.am	49	0.01	0.05	0.108	0.080	0.108	0.074	0.062	0.074
26	semi.par.x1	2d.am	49	0.01	0.15	0.098	0.070	0.098	0.104	0.054	0.104
27	semi.par.x1	2d.am	49	0.04	0.05	0.068	0.058	0.068	0.066	0.034	0.068
28	semi.par.x1	2d.am	49	0.04	0.15	0.066	0.042	0.066	0.076	0.034	0.078
29	semi.par.x1	2d.am	100	0.01	0.05	0.086	0.072	0.086	0.076	0.054	0.078
30	semi.par.x1	2d.am	100	0.01	0.15	0.100	0.066	0.100	0.136	0.092	0.138
31	semi.par.x1	2d.am	100	0.04	0.05	0.088	0.068	0.088	0.072	0.044	0.074
32	semi.par.x1	2d.am	100	0.04	0.15	0.070	0.040	0.070	0.080	0.026	0.080
33	2d.am	2d.sm	49	0.01	0.05	0.152	0.118	0.156	0.136	0.088	0.146
34	2d.am	2d.sm	49	0.01	0.15	0.614	0.424	0.622	0.448	0.148	0.460
35	2d.am	2d.sm	49	0.04	0.05	0.144	0.116	0.146	0.116	0.084	0.118
36	2d.am	2d.sm	49	0.04	0.15	0.104	0.042	0.108	0.080	0.012	0.086
37	2d.am	2d.sm	100	0.01	0.05	0.188	0.146	0.190	0.144	0.104	0.146
38	2d.am	2d.sm	100	0.01	0.15	0.866	0.734	0.868	0.570	0.202	0.572
39	2d.am	2d.sm	100	0.04	0.05	0.162	0.132	0.162	0.122	0.084	0.122
40	2d.am	2d.sm	100	0.04	0.15	0.162	0.088	0.168	0.100	0.020	0.102

**Table B.11.** Definitions of 'Setups' summarised in Tables B.12 - B.17

Setup	Reduced Model	Full Model	h
1	2d.am	2d.sm	0.05
2	2d.am	2d.sm	0.1
3	2d.am	2d.sm	0.15
4	2d.am	2d.sm	0.2
5	2d.am	2d.sm	0.25
6	1d.sm	2d.am	0.05
7	1d.sm	2d.am	0.1
8	1d.sm	2d.am	0.15
9	1d.sm	2d.am	0.2
10	1d.sm	2d.am	0.25
11	1d.sm	semi.par.x1	0.05
12	1d.sm	semi.par.x1	0.1
13	1d.sm	semi.par.x1	0.15
14	1d.sm	semi.par.x1	0.2
15	1d.sm	semi.par.x1	0.25
16	semi.par.x1	2d.am	0.05
17	semi.par.x1	2d.am	0.1
18	semi.par.x1	2d.am	0.15
19	semi.par.x1	2d.am	0.2
20	semi.par.x1	2d.am	0.25
21	st.line	semi.par.x2	0.05
22	st.line	semi.par.x2	0.1
23	st.line	semi.par.x2	0.15
24	st.line	semi.par.x2	0.2
25	st.line	semi.par.x2	0.25



**Table B.12.** Results of simulation study to investigate the size of tests of model comparisons using RSSD and CFV and 5 estimators of  $\sigma^2$ . Design points are configurations of 100 points from a uniform random design. The results listed are the proportion of 500 simulated p-values less than 0.05 using a *Quadratic Form* reference distribution. ‘Setup’ defined in Table B.11

Setup	CFV					RSSD				
	Diff	Wgtd	h.num/4	h.num/2	h.num	Diff	Wgtd	h.num/4	h.num/2	h.num
1	0.020	0.014			0.036	0.014	0.012			0.032
2	0.064	0.044		0.084	0.030	0.068	0.074		0.096	0.036
3	0.102	0.074	0.084	0.082	0.018	0.138	0.126	0.096	0.134	0.012
4	0.094	0.082	0.102	0.052	0.016	0.154	0.138	0.198	0.092	0.008
5	0.086	0.072	0.082	0.030	0.008	0.170	0.144	0.172	0.058	0.006
6	0.016	0.014	0.074	0.066	0.044	0.004	0.004	0.058	0.050	0.034
7	0.092	0.078	0.162	0.148	0.042	0.040	0.034	0.090	0.072	0.016
8	0.196	0.186	0.254	0.192	0.054	0.134	0.128	0.218	0.128	0.020
9	0.248	0.230	0.324	0.186	0.056	0.230	0.222	0.296	0.142	0.028
10	0.280	0.264	0.316	0.160	0.052	0.280	0.270	0.324	0.138	0.040
11	0.024	0.022	0.040	0.044	0.034	0.030	0.036	0.046	0.046	0.040
12	0.064	0.062	0.078	0.080	0.050	0.038	0.038	0.066	0.064	0.024
13	0.092	0.094	0.124	0.104	0.058	0.072	0.074	0.096	0.084	0.032
14	0.112	0.120	0.134	0.098	0.056	0.106	0.112	0.136	0.086	0.038
15	0.112	0.116	0.134	0.106	0.062	0.126	0.134	0.144	0.080	0.034
16	0.038	0.032	0.058	0.058	0.056	0.034	0.026	0.050	0.052	0.050
17	0.080	0.080	0.120	0.108	0.056	0.054	0.044	0.074	0.070	0.030
18	0.134	0.126	0.166	0.140	0.056	0.096	0.092	0.128	0.090	0.026
19	0.176	0.170	0.220	0.130	0.052	0.146	0.130	0.180	0.104	0.034
20	0.214	0.204	0.236	0.120	0.052	0.174	0.168	0.204	0.094	0.030
21	0.054	0.054	0.054	0.052	0.058	0.054	0.054	0.044	0.050	0.046
22	0.042	0.048	0.056	0.054	0.054	0.048	0.058	0.052	0.050	0.054
23	0.046	0.048	0.050	0.046	0.048	0.050	0.058	0.052	0.056	0.056
24	0.044	0.046	0.050	0.044	0.042	0.052	0.056	0.052	0.054	0.058
25	0.044	0.042	0.046	0.048	0.046	0.052	0.048	0.050	0.044	0.044

**Table B.13.** Results of simulation study to investigate the size of tests of model comparisons using RSSD and CFV and 3 estimators of  $\sigma^2$ . Design points are configurations of 100 points from a uniform random design. The results listed are the proportion of 500 simulated p-values less than 0.05 using a  $F$  reference distribution. 'Setup' defined in Table B.11

Setup	CFV			RSSD		
	h.num/4	h.num/2	h.num	h.num/4	h.num/2	h.num
1			0.020			0.010
2		0.030	0.012		0.056	0.016
3	0.040	0.036	0.006	0.058	0.064	0.006
4	0.050	0.026	0.000	0.110	0.040	0.004
5	0.036	0.016	0.000	0.114	0.028	0.002
6	0.050	0.046	0.034	0.030	0.026	0.016
7	0.132	0.120	0.032	0.070	0.056	0.010
8	0.228	0.174	0.046	0.174	0.106	0.014
9	0.290	0.174	0.040	0.256	0.124	0.020
10	0.298	0.140	0.046	0.290	0.122	0.024
11	0.028	0.024	0.020	0.032	0.036	0.028
12	0.050	0.046	0.028	0.042	0.038	0.020
13	0.072	0.058	0.028	0.074	0.060	0.014
14	0.074	0.052	0.024	0.098	0.060	0.022
15	0.070	0.040	0.020	0.110	0.064	0.024
16	0.058	0.060	0.056	0.058	0.056	0.052
17	0.124	0.116	0.058	0.078	0.070	0.030
18	0.184	0.148	0.056	0.136	0.094	0.026
19	0.232	0.136	0.054	0.182	0.118	0.038
20	0.240	0.126	0.056	0.208	0.102	0.034
21	0.044	0.038	0.038	0.036	0.032	0.030
22	0.044	0.048	0.046	0.036	0.044	0.040
23	0.036	0.038	0.038	0.042	0.042	0.042
24	0.038	0.038	0.040	0.046	0.044	0.040
25	0.044	0.040	0.040	0.040	0.040	0.040

**Table B.14.** Results of simulation study to investigate the size of tests of model comparisons using RSSD and CFV and 5 estimators of  $\sigma^2$ . Design points are configurations of 100 points from a uniform random design. The results listed are the proportion of 500 simulated p-values less than 0.05 using a *two moment corrected F* reference distribution. ‘Setup’ defined in Table B.11

Setup	CFV					RSSD				
	Diff	Wgtd	h.num/4	h.num/2	h.num	Diff	Wgtd	h.num/4	h.num/2	h.num
1	0.018	0.008			0.042	0.004	0.002			0.030
2	0.062	0.042		0.086	0.026	0.070	0.060		0.106	0.022
3	0.102	0.078	0.102	0.084	0.018	0.128	0.118	0.152	0.132	0.012
4	0.106	0.086	0.112	0.058	0.016	0.156	0.142	0.200	0.092	0.008
5	0.090	0.074	0.092	0.036	0.008	0.174	0.154	0.184	0.060	0.006
6	0.014	0.014	0.076	0.066	0.046	0.004	0.004	0.060	0.052	0.024
7	0.102	0.084	0.172	0.154	0.042	0.040	0.034	0.092	0.076	0.016
8	0.200	0.190	0.264	0.196	0.062	0.140	0.130	0.222	0.128	0.020
9	0.262	0.248	0.336	0.194	0.058	0.232	0.234	0.300	0.150	0.028
10	0.290	0.288	0.328	0.168	0.056	0.284	0.274	0.336	0.140	0.040
11	0.024	0.022	0.044	0.046	0.036	0.022	0.030	0.048	0.046	0.040
12	0.070	0.066	0.082	0.080	0.050	0.040	0.038	0.070	0.066	0.024
13	0.098	0.104	0.130	0.110	0.058	0.076	0.086	0.100	0.088	0.034
14	0.122	0.126	0.142	0.112	0.060	0.120	0.112	0.138	0.090	0.038
15	0.122	0.124	0.144	0.106	0.064	0.132	0.134	0.148	0.096	0.034
16	0.044	0.036	0.062	0.062	0.060	0.040	0.034	0.058	0.058	0.052
17	0.102	0.088	0.124	0.116	0.058	0.062	0.052	0.080	0.072	0.030
18	0.152	0.136	0.188	0.148	0.058	0.110	0.100	0.136	0.094	0.026
19	0.194	0.186	0.232	0.138	0.054	0.156	0.144	0.182	0.118	0.038
20	0.226	0.220	0.240	0.126	0.056	0.190	0.182	0.208	0.104	0.034
21	0.052	0.052	0.056	0.056	0.058	0.048	0.046	0.048	0.050	0.044
22	0.042	0.054	0.058	0.056	0.056	0.048	0.058	0.052	0.054	0.052
23	0.048	0.050	0.050	0.050	0.048	0.052	0.058	0.052	0.056	0.056
24	0.048	0.052	0.050	0.050	0.048	0.056	0.056	0.052	0.056	0.058
25	0.050	0.052	0.048	0.048	0.048	0.054	0.050	0.050	0.048	0.044

**Table B.15.** Results of simulation study to investigate the size of tests of model comparisons using RSSD and CFV and 5 estimators of  $\sigma^2$ . Design points are configurations of 49 points from a uniform random design. The results listed are the proportion of 500 simulated p-values less than 0.05 using a *Quadratic Form* reference distribution. 'Setup' defined in Table B.11

	CFV					RSSD				
	Diff	Wgtd	h.num/4	h.num/2	h.num	Diff	Wgtd	h.num/4	h.num/2	h.num
1	0.002	0.006			0.000	0.002	0.006			0.000
2	0.032	0.026		0.002	0.028	0.022	0.026		0.002	0.020
3	0.050	0.056		0.066	0.028	0.046	0.050		0.058	0.030
4	0.074	0.070	0.012	0.084	0.030	0.072	0.076	0.006	0.100	0.028
5	0.082	0.090	0.096	0.064	0.026	0.108	0.112	0.088	0.086	0.032
6	0.002	0.002	0.002	0.040	0.040	0.000	0.002	0.002	0.014	0.024
7	0.036	0.034	0.114	0.102	0.036	0.012	0.020	0.060	0.062	0.016
8	0.100	0.086	0.204	0.140	0.032	0.058	0.058	0.156	0.104	0.012
9	0.126	0.110	0.246	0.154	0.034	0.104	0.098	0.228	0.114	0.018
10	0.148	0.144	0.248	0.128	0.040	0.150	0.136	0.276	0.104	0.020
11	0.018	0.020	0.000	0.050	0.046	0.004	0.014	0.000	0.028	0.040
12	0.048	0.052	0.076	0.074	0.054	0.034	0.040	0.054	0.058	0.036
13	0.080	0.090	0.106	0.090	0.058	0.062	0.066	0.108	0.092	0.046
14	0.100	0.096	0.128	0.096	0.064	0.088	0.094	0.126	0.100	0.054
15	0.108	0.112	0.138	0.100	0.070	0.100	0.104	0.140	0.090	0.054
16	0.022	0.018	0.042	0.044	0.046	0.018	0.012	0.042	0.050	0.046
17	0.036	0.034	0.108	0.106	0.030	0.020	0.016	0.064	0.058	0.020
18	0.068	0.060	0.156	0.118	0.036	0.048	0.040	0.108	0.084	0.016
19	0.098	0.100	0.182	0.122	0.030	0.074	0.062	0.148	0.092	0.018
20	0.124	0.120	0.196	0.106	0.032	0.084	0.090	0.162	0.084	0.020
21	0.036	0.058	0.034	0.072	0.064	0.024	0.048	0.034	0.062	0.068
22	0.036	0.044	0.044	0.040	0.038	0.036	0.064	0.058	0.054	0.050
23	0.038	0.038	0.038	0.032	0.030	0.042	0.050	0.044	0.038	0.036
24	0.036	0.040	0.032	0.028	0.028	0.038	0.044	0.036	0.036	0.032
25	0.042	0.050	0.034	0.034	0.032	0.040	0.052	0.034	0.036	0.036

**Table B.16.** Results of simulation study to investigate the size of tests of model comparisons using RSSD and CFV and 3 estimators of  $\sigma^2$ . Design points are configurations of 49 points from a uniform random design. The results listed are the proportion of 500 simulated p-values less than 0.05 using a  $F$  reference distribution. 'Setup' defined in Table B.11

	CFV			RSSD		
	h.num/4	h.num/2	h.num	h.num/4	h.num/2	h.num
1			0.000			0.000
2		0.010	0.010		0.010	0.004
3		0.038	0.022		0.044	0.010
4	0.038	0.044	0.012	0.032	0.048	0.014
5	0.058	0.034	0.010	0.084	0.050	0.010
6	0.012	0.020	0.018	0.010	0.014	0.014
7	0.088	0.088	0.028	0.048	0.058	0.006
8	0.182	0.112	0.026	0.134	0.088	0.006
9	0.220	0.136	0.024	0.210	0.104	0.012
10	0.224	0.106	0.036	0.246	0.086	0.012
11	0.008	0.036	0.024	0.006	0.028	0.016
12	0.046	0.050	0.036	0.044	0.042	0.022
13	0.080	0.070	0.038	0.082	0.066	0.024
14	0.082	0.064	0.034	0.102	0.076	0.036
15	0.078	0.056	0.032	0.106	0.078	0.036
16	0.048	0.046	0.050	0.044	0.054	0.048
17	0.114	0.110	0.044	0.070	0.064	0.020
18	0.166	0.126	0.044	0.112	0.090	0.016
19	0.194	0.132	0.042	0.156	0.094	0.020
20	0.208	0.110	0.042	0.172	0.090	0.020
21	0.034	0.038	0.044	0.050	0.046	0.038
22	0.034	0.032	0.026	0.048	0.040	0.036
23	0.030	0.030	0.028	0.034	0.030	0.026
24	0.030	0.026	0.024	0.028	0.026	0.026
25	0.030	0.028	0.024	0.028	0.028	0.026

**Table B.17.** Results of simulation study to investigate the size of tests of model comparisons using RSSD and CFV and 5 estimators of  $\sigma^2$ . Design points are configurations of 49 points from a uniform random design. The results listed are the proportion of 500 simulated p-values less than 0.05 using a *two moment corrected F* reference distribution. 'Setup' defined in Table B.11

	CFV					RSSD				
	Diff	Wgtd	h.num/4	h.num/2	h.num	Diff	Wgtd	h.num/4	h.num/2	h.num
1	0.002	0.002			0.022	0.000	0.000			0.016
2	0.038	0.026		0.052	0.028	0.022	0.024		0.062	0.016
3	0.054	0.054		0.076	0.028	0.056	0.050		0.100	0.018
4	0.084	0.072	0.110	0.088	0.030	0.096	0.076	0.130	0.100	0.024
5	0.092	0.096	0.126	0.076	0.026	0.120	0.112	0.176	0.086	0.028
6	0.004	0.002	0.040	0.050	0.038	0.002	0.002	0.030	0.036	0.020
7	0.038	0.034	0.120	0.108	0.036	0.014	0.018	0.072	0.062	0.010
8	0.102	0.088	0.218	0.156	0.034	0.064	0.058	0.164	0.108	0.008
9	0.138	0.120	0.268	0.160	0.038	0.104	0.102	0.242	0.116	0.018
10	0.164	0.154	0.264	0.134	0.048	0.158	0.144	0.282	0.110	0.020
11	0.018	0.012	0.046	0.058	0.046	0.004	0.012	0.030	0.046	0.028
12	0.048	0.056	0.084	0.086	0.058	0.040	0.036	0.062	0.058	0.032
13	0.088	0.094	0.114	0.098	0.066	0.062	0.066	0.116	0.094	0.044
14	0.118	0.112	0.140	0.112	0.068	0.094	0.098	0.136	0.106	0.054
15	0.120	0.122	0.146	0.106	0.072	0.106	0.108	0.146	0.100	0.054
16	0.026	0.024	0.056	0.056	0.052	0.028	0.020	0.050	0.058	0.050
17	0.050	0.038	0.120	0.112	0.052	0.026	0.022	0.072	0.066	0.022
18	0.080	0.072	0.170	0.130	0.044	0.054	0.044	0.112	0.090	0.016
19	0.110	0.112	0.194	0.134	0.042	0.078	0.076	0.156	0.096	0.020
20	0.136	0.132	0.212	0.112	0.042	0.102	0.102	0.174	0.090	0.020
21	0.052	0.054	0.066	0.076	0.064	0.036	0.040	0.072	0.066	0.054
22	0.036	0.046	0.046	0.040	0.042	0.046	0.054	0.058	0.054	0.046
23	0.042	0.038	0.038	0.036	0.032	0.044	0.048	0.046	0.040	0.034
24	0.038	0.040	0.032	0.034	0.032	0.040	0.048	0.040	0.038	0.032
25	0.046	0.054	0.042	0.038	0.036	0.042	0.052	0.036	0.038	0.036

**Table B.18.** Results of simulation study to investigate the power of tests of model comparisons using RSSD and CFV. The first row's results are based on a difference based estimator of  $\sigma^2$  and the remaining rows are based on a RSS based estimator of  $\sigma^2$  (no undersmoothing). Design points are configurations from a uniform random design. The results listed are the proportion of 500 simulated p-values less than 0.05 using a *two moment corrected F* reference distribution.

Red. Model	Full Model	n	Sm. Par.	CFV	RSSD
2d.am	2d.sm	100	0.1	0.740	0.798
1d.sm	2d.am	100	0.05	0.868	0.780
1d.sm	2d.am	100	0.1	0.802	0.770
1d.sm	2d.am	100	0.15	0.694	0.680
1d.sm	2d.am	100	0.2	0.634	0.622
1d.sm	2d.am	100	0.25	0.590	0.576
semi.par.x1	2d.am	100	0.05	0.508	0.386
semi.par.x1	2d.am	100	0.1	0.408	0.388
semi.par.x1	2d.am	100	0.15	0.318	0.318
semi.par.x1	2d.am	100	0.2	0.312	0.276
semi.par.x1	2d.am	100	0.25	0.300	0.260
1d.sm	semi.par.x1	100	0.05	0.672	0.672
1d.sm	semi.par.x1	100	0.1	0.562	0.558
1d.sm	semi.par.x1	100	0.15	0.448	0.446
1d.sm	semi.par.x1	100	0.2	0.372	0.372
1d.sm	semi.par.x1	100	0.25	0.344	0.336
st.line	semi.par.x2	100	0.05	0.672	0.672
st.line	semi.par.x2	100	0.1	0.562	0.558
st.line	semi.par.x2	100	0.15	0.448	0.446
st.line	semi.par.x2	100	0.2	0.372	0.372
st.line	semi.par.x2	100	0.25	0.344	0.336
2d.am	2d.sm	49	0.1	0.222	0.210
1d.sm	2d.am	49	0.05	0.430	0.274
1d.sm	2d.am	49	0.1	0.408	0.304
1d.sm	2d.am	49	0.15	0.328	0.276
1d.sm	2d.am	49	0.2	0.306	0.256
1d.sm	2d.am	49	0.25	0.268	0.236
semi.par.x1	2d.am	49	0.05	0.164	0.100
semi.par.x1	2d.am	49	0.1	0.170	0.112
semi.par.x1	2d.am	49	0.15	0.140	0.102
semi.par.x1	2d.am	49	0.2	0.134	0.108
semi.par.x1	2d.am	49	0.25	0.140	0.106
1d.sm	semi.par.x1	49	0.05	0.368	0.348
1d.sm	semi.par.x1	49	0.1	0.312	0.282
1d.sm	semi.par.x1	49	0.15	0.226	0.202
1d.sm	semi.par.x1	49	0.2	0.198	0.172
1d.sm	semi.par.x1	49	0.25	0.186	0.166
st.line	semi.par.x2	49	0.05	0.368	0.348
st.line	semi.par.x2	49	0.1	0.312	0.282
st.line	semi.par.x2	49	0.15	0.226	0.202
st.line	semi.par.x2	49	0.2	0.198	0.172
st.line	semi.par.x2	49	0.25	0.186	0.166

# Bibliography

- Azzalini, A. and Bowman, A. W. (1993). On the use of nonparametric regression for checking linear relationships. *JRSS Series B*, 55:549–557.
- Azzalini, A., Bowman, A. W., and Härdle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika*, 76:1–11.
- Bellman, R. E. (1961). *Adaptive control processes*. Princeton University Press.
- Bock, M. T. (1999). Some issues related to comparisons of nonparametric regression models. In Friedl, H., Berghold, A., and Kauermann, G., editors, *Statistical Modelling*, pages 110–117. 14th International Workshop on Statistical Modelling.
- Bock, M. T., Miller, B. S., and Bowman, A. W. (1999). Assessment of eutrophication in the Firth of Clyde: Analysis of coastal water data from 1982 to 1996. *Marine Pollution Bulletin*, 38(3):222–231.
- Boney, A. D. (1986). Seasonal studies on the phytoplankton and primary production in the inner Firth of Clyde. In *Proceedings of the Royal Society of Edinburgh*, volume 90B, pages 203–222.
- Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford University Press.
- Bowman, A. W. and Young, S. G. (1996). Graphical comparisons of nonparametric curves. *Applied Statistics*, 45:83–98.



- Breiman, L. and Meisel, W. S. (1976). General estimates of the intrinsic variability of data in nonlinear regression models. *JASA*, 71:301–307.
- Buckley, M. J., Eagleson, G. K., and Silverman, B. W. (1988). The estimation of residual variance in nonparametric regression. *Biometrika*, 75:189–199.
- Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Annals of Statistics*, 17:453–555.
- Carroll, R. J. (1982). Adapting for heteroscedasticity in linear models. *The Annals of Statistics*, 11:1224–1223.
- Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *JASA*, 92:477–89.
- Carter, C. K. and Eagleson, G. K. (1992). A comparison of variance estimators in nonparametric regression. *JRSS Series B*, 54:773–780.
- Chambers, J. M. and Hastie, T. J. (1993). *Statistical models in S*. Chapman and Hall, London.
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *JASA*, 74:829–836.
- Cleveland, W. S. (1993). *Visualizing Data*. Hobart Press, Summit, NJ.
- Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *JASA*, 83(403):597–610.
- Cleveland, W. S. and Loader, C. (1996). Smoothing by local regression: principles and methods. *Statistical theory and computational aspects of smoothing* (W. Härdle and M.G. Schimek, eds.), Springer, New York.
- Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall, New York.
- Cox, D. R. (1972). Analysis of survival data. *JRSS Series B*, 34:187–220.

- Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman and Hall.
- CSTT (1994). Comprehensive studies for the purpose of article 6 of Dir 91/271/EEC, the Urban Waste Water Treatment Directive. Comprehensive Studies Task Team, Edinburgh.
- Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge University Press.
- Dette, H. and Munk, A. (1998). A simple goodness-of-fit test for linear models under a random design assumption. *Annals of the Institute of Statistical Mathematics*, 50:253–75.
- Dette, H., Munk, A., and Wagner, T. (1998). Estimating the variance in nonparametric regression - what is a reasonable choice? *JRSS Series B*, 60:751–764.
- Dette, H., Munk, A., and Wagner, T. (1999). A review of variance estimators with extensions to multivariate nonparametric regression models. *Multivariate Design and Sampling (S. Gosh, ed.)*, New York, Dekker: To be published.
- Eagleson, G. K. (1989). Curve estimation - whatever happened to the variance? *Proc. 47th Sess. Int. Statist. Inst.* 535–551, pages 535–551.
- Engle, R., Granger, C., Rice, J., and Weiss, A. (1986). Nonparametric estimates of the relation between weather and electricity sales. *JASA*, 81:310–320.
- Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York and Basel.
- Fan, J. (1992). Design-adaptive nonparametric regression. *JASA*, 87:196–216.
- Fan, J., Farman, M., and Gijbels, I. (1998a). Local maximum likelihood estimation and inference. *JRSS Series B*, 60:591–608.
- Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman and Hall, London.

- Fan, J., Härdle, W., and Mammen, E. (1998b). Direct estimation of low-dimensional components in additive models. *Annals of Statistics*, 26:943–971.
- Firth, D., Glossup, J., and Hinkley, D. V. (1991). Model checking with nonparametric curves. *Biometrika*, 78:245–252.
- Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, 19:1–141.
- Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *JASA*, 76:817–823.
- Gasser, T. and Müller, H. G. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation* (T. Gasser and M. Rosenblatt, eds.), Springer-Verlag, Berlin:23–68.
- Gasser, T. and Seifert, B. (1994). An answer to some points raised by Cleveland and Loader. Available at [www.unizh.ch/biostat/People/gasser/](http://www.unizh.ch/biostat/People/gasser/).
- Gasser, T., Sroka, L., and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, 73:625–33.
- Gjertson, D. W. (1991). Survival trends in long term first cadaver donor kidney transplants. *Teraski, P.I. Ed. Clinical Transplants 1991. Los Angeles UCLA. Tissue Typing Authority*, pages 225–235.
- Glibert, P. M. (1988). Primary productivity and pelagic nitrogen cycling. *Nitrogen Cycling in Coastal Marine Environments* (T. H. Blackburn and J. Sorensen, eds.), Wiley, Chichester.
- Green, P., Jennison, C., and Seheult, A. (1985). Analysis of field experiments by least squares smoothing. *JRSS Series B*, 47:299–315.
- Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. Chapman and Hall, London.

- Hall, B. M., Tiller, D. J., and Hardie, I. (1988). Comparison of three immunosuppressive regimens in cadaver renal transplantation: long-term cyclosporin, short term cyclosporin followed by azathioprine prednisolone without cyclosporin. *N. Engl. J. Med.*
- Hall, P. and Carroll, R. J. (1989). Variance function estimation in regression: the effect of estimating the mean. *JRSS Series B*, 51:3–14.
- Hall, P. and Hart, J. D. (1990). Bootstrap test for difference between means in nonparametric regression. *JASA*, 85:1039–49.
- Hall, P., Kay, J. W., and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77:521–8.
- Hall, P., Kay, J. W., and Titterington, D. M. (1991). On estimation of noise variation in signal-processing. *Advances in applied probability*, 23:476–495.
- Hall, P. and Marron, J. S. (1990). On variance estimation in nonparametric regression. *Biometrika*, 77:415–9.
- Härdle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, 21:157–78.
- Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, 21:1926–1947.
- Härdle, W. and Scott, D. W. (1992). Smoothing by weighted averaging of rounded points. *Computational Statistics*, 7:97–128.
- Hart, J. D. (1997). *Nonparametric smoothing and lack-of-fit test*. Springer-Verlag, New York.
- Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.
- Herrmann, E., Wand, M. P., Engel, J., and Gasser, T. (1995). A bandwidth selector for bivariate kernel regression. *JRSS Series B*, 57:171–180.

- Hollander, A. A. M. J., van Saase, J. L. C. M., and Kootte, A. M. M. (1995). Beneficial effects of conversion from cyclosporin to asathioprine after kidney transplantation. *Lancet*.
- Hurvich, C. M., Simonoff, J. S., and Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike Information Criterion. *JRSS Series B*, 60:271–293.
- Johnson, N. L. and Kotz, S. (1972). *Distributions in Statistics: Continuous Univariate Distributions, Vol. II*. Wiley, New York.
- Kester, D. A. (1975). Dissolved gases other than CO<sub>2</sub>. *Chemical Oceanography* (J. P. Riley and G. Skirrow, eds.), Academic Press, London.
- King, E. C., Hart, J. D., and Wehrly, T. E. (1991). Testing the equality of two regression curves using linear smoothers. *Statistical Probability Letters*, 12:239–47.
- Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82:93–100.
- MacPhee, I. A. M., Bradley, J. A., and Briggs, J. D. (1998). Longterm outcome of a prospective randomised trial of conversion from cyclosporin to azathioprine treatment one year after renal transplantation. *Transplantation*, 66:1186–92.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, 2nd edn*. Chapman and Hall, London.
- Mickey, R., Cho, Y. W., and Carnahan, E. (1990). Long-term graft survival. *Teraski, P.I. Ed. Clinical Transplants 1990. Los Angeles UCLA. Tissue Typing Authority*, pages 385–396.
- Morris, P. J. (1994). *Kidney Transplantation*. W.B. Saunders, London.
- Morris, P. J., Chapman, J. R., and Allen, R. D. M. (1987). Cyclosporin conversion versus conventional immunosuppression: long-term follow-up and histological evaluation. *Lancet*.

- Müller, H. G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics*, 15:610–625.
- Myers, B. D., Newton, L., and Oyer, P. (1991). The case against the indefinite use of cyclosporin. *Transplant. Proc.*
- Nadaraya, E. A. (1964). On estimating regression. *Theory of probability and its applications*, 10:186–190.
- Neumann, M. H. (1994). Full data driven nonparametric variance estimators. *Statistics*, 25:199–212.
- Opsomer, J. D. (1999). On the properties of backfitting estimators. *Journal of Multivariate Analysis*. To appear.
- Opsomer, J. D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics*, 25:186–211.
- Opsomer, J. D. and Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modelling. *Journal of Computational and Graphical Statistics*, To appear.
- Raz, J. (1990). Testing for no effect when estimating a smooth function by nonparametric regression: a randomisation approach. *JASA*, 85:132–138.
- Redfield, A. C., Ketchum, B. H., and Richards, F. A. (1963). *The Sea: Ideas and Observations on Progress in the Study of the Seas*, volume II, chapter 2, pages 26–77. Interscience Publishers, New York.
- Reinsch, C. (1967). Smoothing by spline functions. *Numer. Math.*, 10:177–183.
- Rice, J. (1984). Bandwidth choice for nonparametric kernel regression. *Annals of Statistics*, 12:1215–1230.
- Rice, J. (1986). Convergence rates for partially splined models. *Statistical Probability Letters*, 4:203–208.
- Ripley, B. (1981). *Spatial Statistics*. Wiley, New York.

- Rowe, P. A., Watson, M. A., McMillan, M. A., Briggs, J. D., Rodger, R. S. C., and Junor, B. J. R. (1994). Prospective randomised study of the influence of cyclosporin on long-term renal allograft function. *Transplant. Proc.*
- Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *JASA*, 90:1257–1270.
- Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, 22:1346–1370.
- Ruppert, D., Wand, M. P., Holst, U., and Hossjer, O. (1997). Local polynomial variance-function estimation. *Technometrics*, 39:262–273.
- Salomon, D. R. (1992). Cyclosporine nephrotoxicity and long term renal transplantation. *Transplantation Reviews*, pages 10–19.
- Seber, G. A. F. (1977). *Linear Regression Analysis*. John Wiley & Sons.
- Seifert, B., Gasser, T., and Wolf, A. (1993). Nonparametric estimation of residual variance revisited. *Biometrika*, 80:373–83.
- Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *JASA*, 89(426):501–511.
- Severini, T. A. and Wong, W. H. (1992). Generalized profile likelihood and conditionally parametric models. *Annals of Statistics*, 20:1768–1802.
- Shiau, J., Wahba, G., and Johnson, D. R. (1986). Partial spline models for the inclusion of tropopause and frontal boundary information in otherwise smooth two and three dimensional objective analysis. *J. Atmos. Ocean. Technol.*, 3:714–725.
- Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist.*, 12:898–916.
- Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *JRSS Series B*, 47:1–52.

- Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer-Verlag, New York.
- Simonoff, J. S. and Tsai, C. L. (1999). Semiparametric and additive model selection using an improved AIC criterion. *Journal of Computational and Graphical Statistics*, 8.
- Simpson, J. H. and Rippeth, T. P. (1993). The Clyde Sea: a model of the seasonal cycle of stratification and mixing. *Estuarine, Coastal and Shelf Science*, 37:129–144.
- Speckman, P. L. (1988). Kernel smoothing in partial linear models. *JRSS Series B*, 50:413–36.
- UWWTD (1991). Council directive concerning urban waste water treatment (91/271/EEC). Official journal No. L 135/40. *Council of European Communities*, Brussels.
- Wahba, G. (1977). A survey of some smoothing problems and the method of generalized cross-validation for solving them. *Applications in Statistics (P.R. Krisnaiah, ed.)*, North Holland, Amsterdam.
- Wahba, G. (1978). Improper priors, spline smoothing and the problems of guarding against model errors in regression. *JRSS Series B*, 40:364–372.
- Wahba, G. (1983). Bayesian confidence intervals for the cross-validated smoothing spline. *JRSS Series B*, 45:133–150.
- Wahba, G. (1984). Partial spline models for the semiparametric estimation of functions of several variables. *Analyses for Time Series, Japan-US Joint Sem.*, Tokyo: Institute of Statistical Mathematics.
- Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.
- Watson, G. S. (1964). Smooth regression analysis. *Sankhya, Series A*, 26:359–372.



- Whittaker, E. (1923). On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41:63-75.
- Woo, Y. M., Jardine, A. G., and Clark, A. F. (1999). Early graft function and patient survival following cadaver renal transplantation. *Kidney International*, 55:692-9.
- Wright, E. M. (1995). *Nonparametric methods for the exploration and analysis of survival data*. PhD thesis, Dept. of Statistics, University of Glasgow.
- Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *JRSS Series B*, 60:797-811.
- Young, S. G. and Bowman, A. W. (1995). Non-parametric analysis of covariance. *Biometrics*, 51:920-931.

